Universität Stuttgart
Institut für Systemtheorie und Bildschirmtechnik
Lehrstuhl für Systemtheorie und Signalverarbeitung
Professor Dr.-Ing. B. Yang

Diplomarbeit D1086

Emotion Recognition from Speech Signals and Perception of Music

Author: Mélanie Fernández Pradier
Matr.-Nr.: 2539362

Date of work begin: 03.08.2010
Date of submission: 03.02.2011
Supervisor: Prof. Dr.-Ing. Bin Yang
Dipl.-Ing. Fabian Schmieder

Keywords: Emotion Recognition - Music - Speech - Emotions - Feature Generation - Musical Features

This thesis deals with emotion recognition from speech signals. The feature extraction step shall be improved by looking at the perception of music. In music theory, different pitch intervals (consonant, dissonant) and chords are believed to invoke different feelings in listeners. The question is whether there is a similar mechanism between perception of music and perception of emotional speech. Our research will follow three stages. First, the relationship between speech and music at segmental and supra-segmental levels will be analyzed. Secondly, the encoding of emotions through music shall be investigated. In the third stage, a description of the most common features used for emotion recognition from speech will be provided. We will additionally derive new high-level musical features, which will lead us to an improvement of the recognition rate for the basic spoken emotions.

Contents

List of Figures

1. Introduction

1.1. Motivation

> "*Emotion, which is suffering, ceases to be suffering as soon as we form a clear and precise picture of it.*" - Baruch Spinoza

Strong impact of emotions

An emotion is a complex psycho-physiological short-time experience resulting from the interaction of biochemical (internal) and environmental (external) factors [48]. It is interesting to note that the word "eMOTion" (from Latin: *ex-movere* or moving out) shares the same root as the word "MOTivation". Feelings have indeed a very strong impact in our decisions; it is therefore not surprising for the brain to contain so many emotion-related systems. Recent studies have shown that emotions can have memory-enhancing effects in our mind: emotional events actually tend to be recalled more often and with more clarity and details than neutral events.

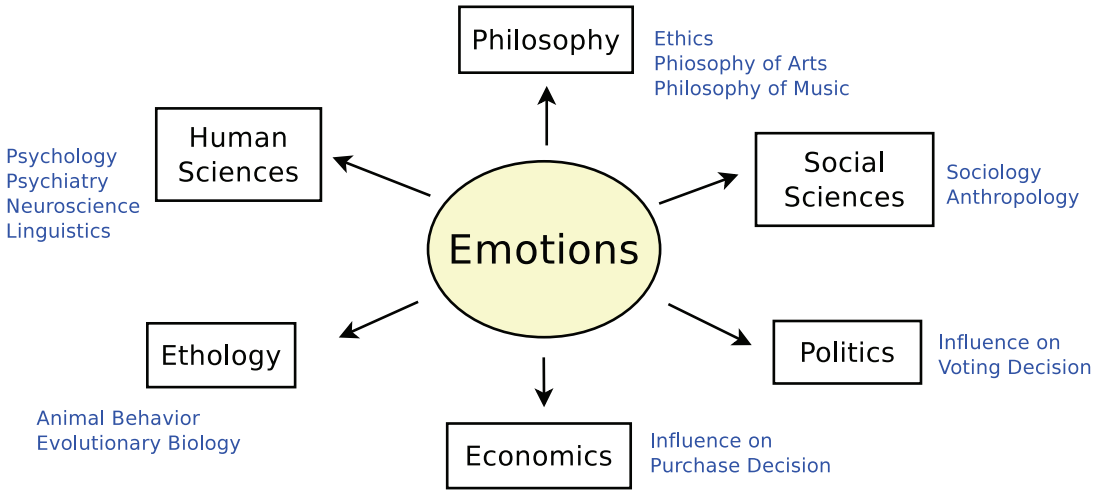


Figure 1.1.: Fields interested in the study of emotions

Throughout History, several scientists have shown interest in the study of emotions [96][71][29]. Darwin states that emotions can be associated with action patterns (see Table 1.1) resulting from natural selection. They are very useful to alert the organism when confronted with important situations. As an example, experiencing fear in front

Stimulus	Cognition	Emotion	Behavior	Effect
threat	"danger"	fear	escape	safety
obstacle	"enemy"	anger	attack	destroy obstacle
gain of valued object	"possess"	happiness	retain or repeat	gain resources
loss of valued object	"abandonment"	sadness	cry	reattach to lost object
member of one's group	"friend"	acceptance	groom	mutual support
unpalatable object	"poison"	disgust	vomit	eject poison
new territory	"examine"	expectation	map	knowledge of territory
unexpected event	"What is it?"	surprise	stop	gain time to orient

Table 1.1.: Association of emotions with action patterns

of a tiger and therefore running away clearly presents an evolutionary advantage for the preservation of the species.

Emotions are essential for the human being to survive, make decisions and preserve his well-being; they influence cognition, perception, learning and communication. This is why very different fields like computer science, psychology or neuroscience are joining their investigative efforts to develop devices that recognize, interpret and process human affects. This thesis belongs to this research wave, focusing on the emotional content of the voice.

Universality of emotion recognition

An interesting question to ask is whether emotional states can be recognized universally or not. Culture and society have a considerable weight on the expression of emotions. This, together with the inherent subjectivity among individuals, can make us wonder about the existence of universal emotions.

If we consider Darwin's theory of evolution, emotions find their root in biology and therefore can be to some extent considered as universals [2]. Several studies have indeed shown evidence for certain universal attributes for both speech [3][80] and music [28][37], not only among individuals of the same culture, but also across cultures. Scherer and Banse, for instance, performed an experiment in which humans had to distinguish between 14 emotions [3]. Their conclusion was that listeners are able to infer emotions from vocal cues, with an average recognition rate 4 or 5 times above chance. He

also observed that not all the emotions were equally well identified. For example, identification of disgust through the voice is generally not very accurate; humans express this emotion in a better way through facial or body expression.

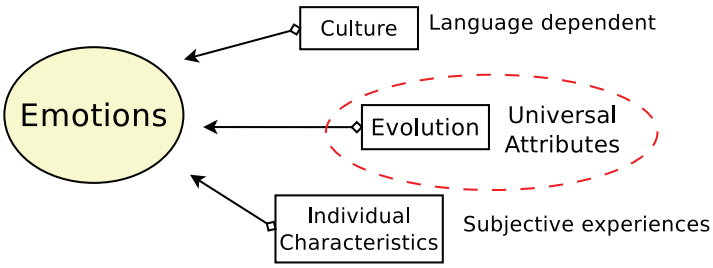


Figure 1.2.: Emotion as the product of evolution, culture and individual traits

Representation of emotions

Since emotions are the result of highly subjective experiences, it is hard to find uniform rules or universal models to represent them. Roughly speaking, there exist two tendencies in the psychological literature, depending on whether we consider emotions discrete or continuous.

The discrete approach consists in designating basic emotional classes. Ekman for instance defined seven basic emotions: happiness, sadness, anger, anxiety, boredom, disgust and neutral [21]. More complex emotions can be seen as mixtures of the basic ones[1]. Another example is the BEEV (Basic English Emotional Vocabulary), which consists of 40 discrete words for automatic emotion recognition [15].

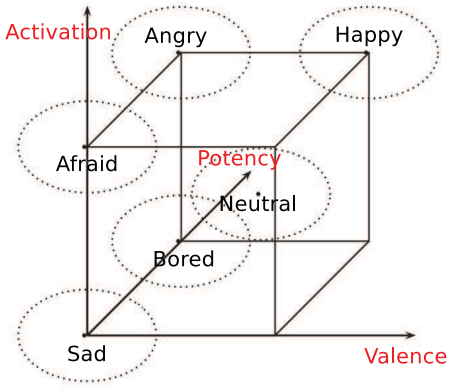


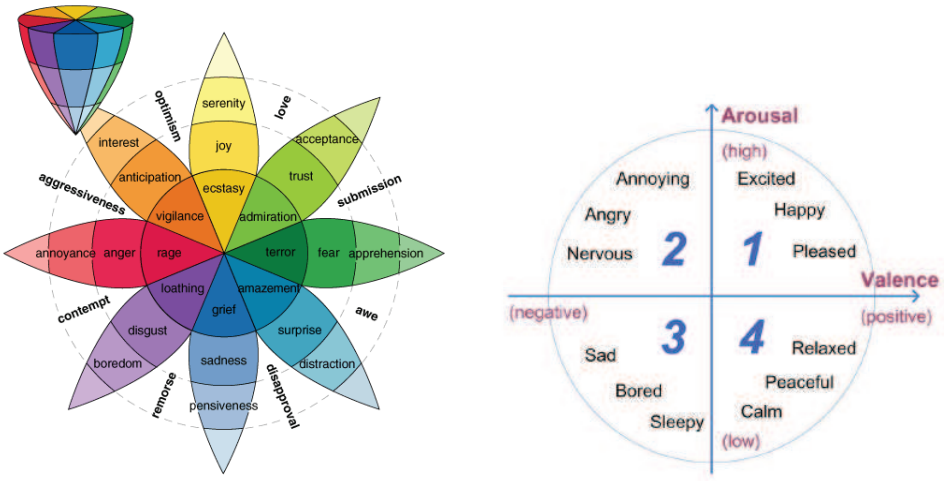
Figure 1.3.: Three-dimensional emotion space and 6 basic emotions

The continuous approach on the other hand consists in defining an N-dimensional emotional space. The most famous one is the three-dimensional model proposed by Schlosberg [81]. Each emotion can be expressed as a linear combination of valence (or

[1]This theory is the so-called "palette theory" of Descartes.

evaluation), arousal (or activation) and potency (or power). Valence defines how positive or negative an emotion is; arousal measures the degree of excitement or involvement of the individual in the emotional state; potency accounts for the strength of the emotion[2].

Both previous approaches can be combined by placing discrete emotions in a continuous space. Figure 1.3 shows the location of six common basic emotions in the Schlosberg space. Other examples are Thayer's valence-arousal representation or Plutchik's emotional wheel [15] (see Figure 1.4).



(a) Plutchik's emotional wheel (b) Thayer's arousal-valence emotion plane

Figure 1.4.: Different representations of emotional states

Emotion recognition from speech

Speech can be seen as a two-channel mechanism, involving not only actual meaning of the communication but also several prosodic nuances. The *linguistic channel* deals with the actual information inferred by words ("What is said") whereas the *paralinguistic channel* gives additional information about the speaker ("How it is said"), namely his emotional state. The linguistic channel was the main focus for research in the past, but scientists have recently become more and more interested in this second implicit channel [99][78]. The increase of computational studies about emotion in speech[3] has given birth to an extensive range of interesting applications [15][14][55] (see Table 1.2).

Perception of music

Music is frequently referred to as the "language of emotion". It has a clear effect on our mood and is sometimes even used in the treatment of affective disorders. Furthermore,

[2] Arousal and potency should not be mixed up. An emotion like boredom has high potency but low arousal.

[3] Examples of well-known studies are the ASSESS system or the Banse and Scherer's System.

Applications	Description
Support to Syntactic Channel	Help to resolve linguistic ambiguities in Automatic Speech Recognition ASR (ex: sarcasm, irony)
Augmenting Human Judgment	Lie detection Clinical diagnoses (detection of schizophrenia, depressions)
Human-Computer Interaction	Improvement of Speech Synthesis: – Generation of convincing emotional speech – Signalization of the stage of transaction – Convergence of vocal parameters between speakers
Tutoring	For computer-based learning, being able to detect Boredom and adapt the teaching methods if required
Avoidance	Detect angry customers in call centers
Entertainment	Electronics with react differently depending on the gathered emotional state of the user

Table 1.2.: Applications of emotion recognition from speech signals

it seems to have the potential to initiate or reinforce social bonding among individuals in a group (for example, in a party or in a concert). Why are humans actually able to experience strong emotions by interacting with music? Even if this process has not been completely understood yet, the analysis of musical perception might help us to better understand emotions [34][33][85][57].

1.2. Aim of the thesis

The final objective of this project is the improvement of emotion recognition from speech. Generally speaking, the possible research directions in this field are the following:

- Modelization: find a better model of emotions (more emotional labels, clearer relationships between emotions, look for relative measures in relation to the neutral state).
- Feature Extraction: look for new features that would better capture the essence of emotional states.
- Classification: investigate algorithms that would perform the better classification for a given set of features.

– Environment: use bigger and more complex databases (natural, with verbal content, wide range of speakers, languages).

This thesis concerns mostly the feature extraction task. In the past, so called prosodic standard features were extracted from the speech signal, directly based on pitch, energy, and duration. In our case we will extract features inspired on music theory. Our aim is to investigate whether there is a similar mechanism between the perception of music and the perception of emotional speech.

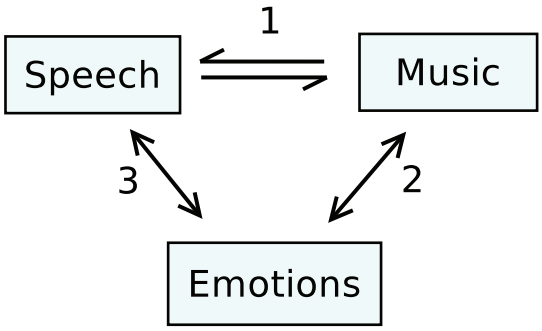


Figure 1.5.: Aim of the thesis

Figure 1.5 shows the three principal concerns of the thesis, namely:

1. Is there any link between music and speech?
2. How is the emotional message transmitted through music?
3. How do we recognize emotions from speech signals? Is it possible to apply music theory to extract new universal features?

An additional remark concerning emotion recognition is that anger and happiness are difficult to distinguish, even if this task does not seem very difficult to humans. If we take a look at the three-dimensional emotional space (Figure 1.3), it can be seen that anger and happiness only differ in the valence dimension. So far, no distinctive acoustic correlates for valence have been found. There is nevertheless an intuition that a musical analysis of speech may help to solve this problem [53].

Encountered difficulties and restrictions

In this thesis, we will restrict ourselves to the analysis of acted emotions. Discrete emotional classes will be used for classification. There are however several factors that severely complicate the task of emotion recognition, but that unfortunately cannot be avoided. The most important ones are listed below:

– Ambiguity (ex: lowered eyebrows can mean both anger or despair)

– Deception (people sometimes hide their real emotions by acting)

- Influence of society (so-called *display rules*, ex: a burst of anger should not be fully expressed)
- Influence of language (ex: Italian often considered more melodious than Russian)
- Interactions within the paralinguistic channel: the same channel is used for multiple functions. Some of them (like emphasis) are related to the linguistic channel. Concerning the features, the role of acoustical changes depends dramatically on the temporal location in the utterance.

1.3. Outline

This thesis is organized as follows: first of all, an inter-disciplinary review concerning emotion, voice and music will be presented in Chapters 2 and 3. The first one describes different studies supporting the existence of a strong connexion between speech and music. Chapter 3 will then investigate why and how music can transmit affects. Musical concepts essential to our study (like dissonance, tension or harmony) will be explained herein.

The next Chapters are more directly concerned with the problem of emotion recognition in speech. Chapter 4 exposes important characteristics of speech, traditionally used features as well as important algorithms for speech processing. Afterward, the implementation of diverse music-based features will be described in Chapter 5.

Chapter 6 gathers the most interesting simulations that were run in MATLAB, mostly on a German database. Our simulations prove that the usage of additional musical cues can improve our recognition rate. Finally, the conclusion of our analysis and further research topics will be brought in the final Chapter.

2. Link between Emotional Speech and Music Perception

"*Music is the literature of the heart; it commences where speech ends.*" - Alphonse de Lamartine

2.1. Interesting facts

Music-language relationships have always been of interest for the scientific community [65][23][41]. Plato already asserted that the power of certain musical modes to uplift the spirit stemmed from their resemblance to the sounds of noble speech. Darwin on the other hand affirmed that both language and music have a common origin in evolution, and that an intermediate form of communication between speech and music may have been at the origin of our species' communicative abilities [19][52].

It is interesting to notice that both speech and music appear in every single human society. Let us take for instance a small tribe from the Brazilian Amazon called the Pirahã. This tribe have no terms for counting nor colors; they do not have creational myths, and almost no artistic expression (craftwork or drawings are practically nonexistent). Yet, they have music in abundance in the form of songs.

Not less surprising, the combination of music and language seems to be exclusive to our species. Although some animals have music (like whales) or language (like prairie dogs), human beings seem to be the only ones to have distinctly both. An exception to this assumption is the case of parrots: these animals are able to both sing and talk. Curiously, they are the only animals[1] that can dance on tact with music (see Figure 2.1). This fact might be related to their ability to mimic sounds that they hear [68].

Figure 2.1.: Snowball, the dancing cockatoo

[1]Some elephant species from Africa are also sensitive to the beat, but it is still being investigated.

2.2. Comparison between speech and music

Let us explore what the commonalities and dissimilarities between music and language are; these have been summarized in Table 2.1. The first and most evident difference concerns the segmental level: musical units are based on pitch contrast whereas speech units are based on timbral variations (modifications of the acoustic wave shape). Furthermore music organizes pitch and rhythm in ways that speech cannot (there is a vertical dimension called harmony). It also appears to have much deeper power over our emotions compared to ordinary speech.

From his side, language is more specific due to its semantic meaning, and dispose of grammatical categories which are otherwise absent in music. Finally there are some brain damage conditions which sometimes affect one domain but spares the other. We call these Aphasia (language disorder with music perception unaffected) and Amusia (musical disorder mainly to process pitch, but having intact communication skills). All these observations seem to point out that music and speech have little in common.

Speech	**Music**
Timbral contrast	Pitch contrast
Semantic meaning More complex syntax	Verticality: harmony Stronger emotional power
Aphasia	Amusia
RHYTHM: systematic patterns of timing, accent, grouping MELODY: structured patterns of pitch over time SYNTAX: discrete elements + principles of combination AFFECTS: extraction of emotional meanings from acoustic signals – Similar processing steps: – Encoding: formation of learned sound categories – Feature extraction: statistical regularities from rhythmic and melodic sequences – Integration: incoming elements into syntactic structures	

Table 2.1.: Comparison between speech and music

Despite of these differences, there are some evidences of an underlying connexion between both domains. For example the mechanism to identify different melodic lines is very similar to how we manage to filter a conversation in a crowdy place (the so-called cocktail party effect). Also there exist some therapies against Aphasia in which speech skills are stimulated through singing.

Both domains share indeed the auditory system and some neuronal processing. Even if each one has its specialized representation set (pitch intervals in music or words in language), some perceptual mechanisms are shared. In both cases, a continuous signal is

received and interpreted in terms of distinct discrete categories. Rhythmic and melodic sequences (such as words and musical tones) are then identified and finally integrated into syntactic structures.

Last but no least, emotional meanings can be transmitted either through speech or music. But are the conveyed emotions of the same nature? Figure 2.2 shows a schematic drawing of the emotion types that music and language are able to transmit. Music can provoke a wide range of sensations (sometimes referred to as SEM or Strong Experience of Music) which often result in chills or goose bumps. These can be analogous to basic everyday's emotions (ex: happiness, anger or fear) or purely aesthetic ones (ex: admiration of beauty, sublime or wonder). On the other hand, speech allows the delivery of not only basic emotions, but also more complex ones involving external factors like jealousy, pride or regret.

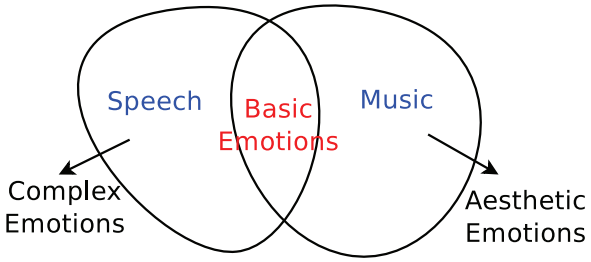


Figure 2.2.: Nature of emotions in speech and music

2.3. The melody of speech

In linguistics, we call *prosody* the ensemble of rhythm, stress, and intonation of speech. Musical melody and speech prosody seems to be connected in some way, but it has to be conceded that true melodies and triadic harmonies are quite infrequent in normal speech. Why isn't speech more melodic in general? The answer can be found in the structure of our brain [72][40]. It is well known that speech and music processing activate different areas in our brain (see Figure 2.3). The left hemisphere deals with the language whereas the right hemisphere is concerned with arts and music. The control of speech melody resides in the inter-hemispheric communication, which will be more or less opened depending if we are more or less emotional.

Even if music-like harmonies do not frequently appear in normal speech, pitch variations in speech are always present, usually in the range of a musical 5th. Why isn't neutral speech flat at all? This answer resides in the other functions of intonation: apart from emotional information, prosody serves other linguistic functions like catching the attention of the listener or putting more or less emphasis in the sentence. It also encodes for the nature of the utterance (whether it is a statement, a question, or a command), irony, sarcasm or contrast.

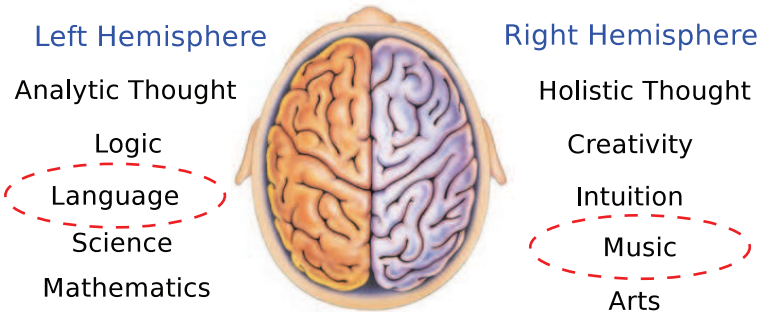
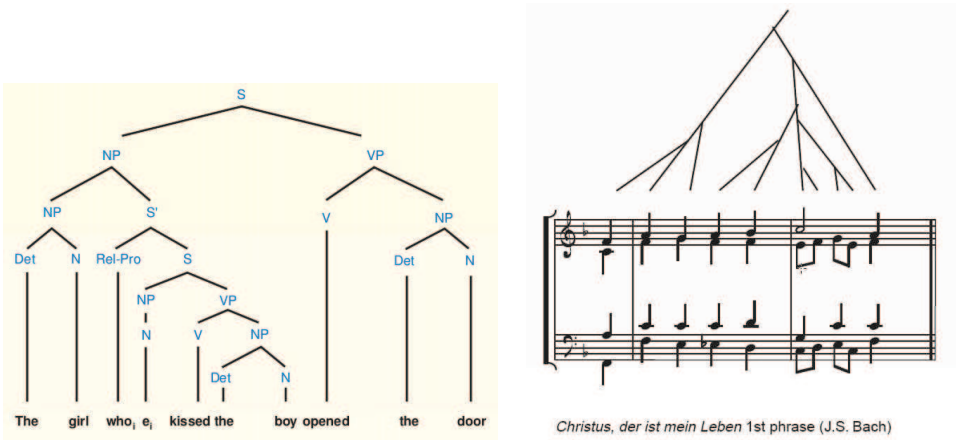


Figure 2.3.: Lateralization of the brain

2.4. Link in neuroscience: shared syntactic integration

Neurologists have recently shown a strong interest in the brain electrical activity during prosodic and melodic processing [5][44][76][39]. Several experiments manipulating the F0 have revealed that both types of processing generate similar electrophysiological signals and are therefore closely related [64][63]. Yet, we have previously seen that there exist some neuropsychological disorders called Aphasia and Amusia that affect only one of the two systems. There is thus an apparent contradiction between neuroimaging (same areas activated) and neuropsychology (Aphasia and Amusia).

An explanation to this problem can be found if we consider syntactic representation and syntactic processing separately. Although speech and music may have their own representation systems, the syntactic processing unit may be unique. This is called the Shared Syntactic Integration Resource Hypothesis (SSIRH). Evidences supporting this conjecture can be found in [66][22], where comparative syntactic analysis for both music and speech are provided (see Figure 2.4).



(a) Speech sentence (b) Musical utterance

Figure 2.4.: Analogous syntactic analysis

Syntax in music refers to the pattern of tension and resolution in time: it is possible to build hierarchical syntactic trees within the Tonal Pitch Space (TPS) theory. The integration cost in syntactic analysis depends in both cases on the distance between the related words or chords (for more details, please refer to the Dependency Locality Theory for language and Tonal Pitch Space Theory for music).

An interesting experiment consisted in measuring the syntactic integration costs when processing simultaneously both speech and music. The observed result was an increment in processing time: there is indeed a super-additive processing difficulty due to competition for limited resources. Hence language and music seem to share the same syntactic processing unit.

2.5. Link in linguistics: rhythm and melody

Another surprising speech-music relational study compares the rhythm and melody of speech and music in the case of British English and French [69][67]. The hypothesis of this experiment is that prosody of a culture's native language influences rhythm and melody of its instrumental music. In other words, French and English composers would tend to compose melodies which resemble respectively to French or English languages.

By rhythm in speech, we refer to a systematic temporal and accentual patterning of sound. As a general rule, language tends to be rhythmically divided into equal portions of time. This division is called isochrony, and can be of three different types:

- *stress-timed*: stress uniformly distributed along time
- *syllable-timed*: syllables perceived as roughly taking up the same amount of time
- *mora-timed*: equal duration of each mora[2] (or unit of sound)

Syllable-timed	Stress-timed	Mora-timed
French Spanish Finnish Slovene	English German Swedish Norwegian Dutch Portuguese Russian	Japanese Gilbertese Luganda

Table 2.2.: Examples of syllable-, stress- and mora-timed languages

[2]In Japanese for example, the words Tōkyō (to-o-kyo-o), Ōsaka (o-o-sa-ka), or Nagasaki (na-ga-sa-ki) all have four moras, even though they have two, three, and four syllables, respectively.

In order to be able to compare speech with music, we will focus on the vowels information. Speech melody shall be represented by a stylized version of the fundamental frequency F0. The measures used for the analysis of rhythm and melody are respectively:

normalized Pairwise Variability Index (nPVI): Measure of the durational contrast between successive elements in a sequence. We should notice that this index is distinct from measures of overall variability (such as standard deviation) in that it is sensitive to the order of the sequence. In our case, we measure the nPVI for vowels durations, which is in stress-timed languages higher than in syllable-time languages, due to the greater degree of vowel reduction. Indeed, the nPVI for English is higher than for French.

$$\text{nPVI} = \frac{100}{m-1} \times \sum \left| \frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}} \right| \tag{2.1}$$

Melodic Interval Variability: Measure of the normalized standard deviation of pitch intervals. This index tries to capture the spread of pitches around the mean pitch. It evaluates whether steps between successive pitches tend to be more uniform or variable in size.

$$\text{MIV} = 100 \times \frac{\text{std (intervals)}}{\text{mean (intervals)}} \tag{2.2}$$

As a final remark, the instrumental music used in this experiment was selected for all English and French composers who were born in the 1800s and died in the 1900s[3]. This era is appropriate for it is recognized as a time of strong musical nationalism.

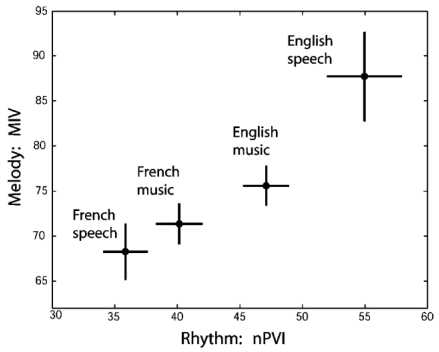


Figure 2.5.: Musico-language study for English and French

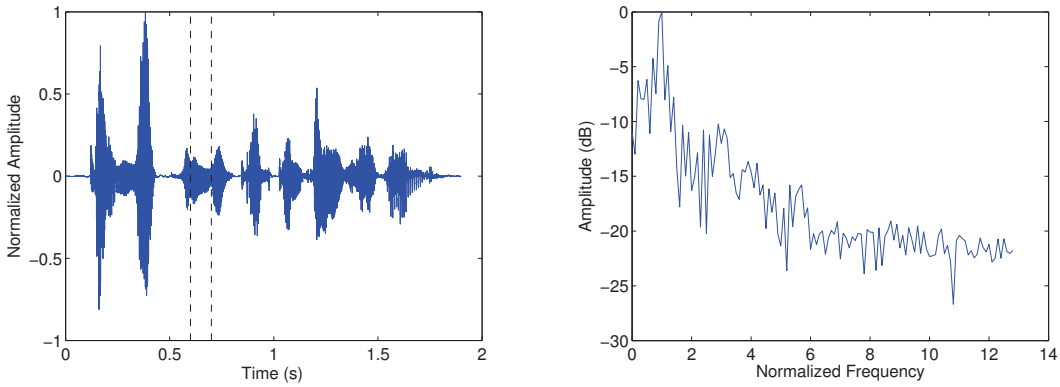
The study concludes with the result of Figure 2.5. In the nPVI-MIV space, it can be observed that French and English music are close to their respective languages. Although not conclusive, this experiment supports the hypothesis that some prosody characteristics of speech are reflected in their respective instrumental music.

[3]Composers where drawn from a musico-logical sourcebook for instrumental music: "A Dictionary of Musical Themes", Barlow and Morgenstern, 1983

2.6. Link in statistics: musical universals

In this section, it will be shown that some musical characteristics can be explained by the nature of our auditory system [87]. The statistical analysis of human speech structure can actually help to understand musical universals like dissonance, chromatic scale or commonly-used intervals. In music theory, intervals like the octave, the fifth or the fourth possess interesting properties that make them stand out from the others.

The procedure in [87] consisted in selecting a huge amount of small 100ms blocks from several speech sentences and compute the local FFT for each of them (see Figure 2.6). A normalization shall then be performed in both amplitude and frequency axis. The frequency values are normalized in respect to the frequency F_m corresponding to the maximum amplitude A_m in the block, and the amplitude axis in respect to this value A_m. F_m often corresponds to the fundamental frequency. In order to remove silent segments, a threshold value of $0.1 \times A_{max}$ for the amplitude shall be used, where A_{max} refers to the maximum amplitude of the whole speech sentence. Finally, an average spectrum from all the FFTs was computed, as it can be seen in Figure 2.7.



(a) Selection of 100ms blocks

(b) Computation of local FFT for each block

Figure 2.6.: Statistical analysis of the human speech structure

Surprisingly, this spectrum presents peaks at some specific frequency ratios, independently of the language or genre of the speakers. The result of this experiment is very impressing: these peaks correspond to specific well-known musical intervals and their distinctive location is apparently universal. The likelihood of different amplitude-frequency combinations is thus related to musical scale structures. The following conclusions can be drawn:

- Why does the chromatic scale usually appear in music? This division in 12 pitch intervals corresponds to the amplitude maxima in the normalized spectrum of speech sounds.
- Why is there usually a preferred subset of tones, namely the Diatonic or Pentatonic scales? These tones correspond to the peaks with greatest concentrations of power.

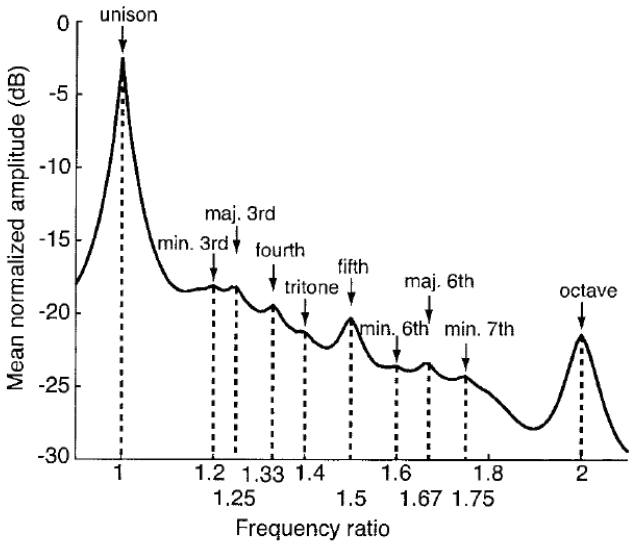


Figure 2.7.: Statistics of speech structure predicts musical universals

– Finally, why do some intervals sound more consonant than others? The consonance ordering follows the ordering of peaks in the spectrum, from the highest to the lowest ones.

2.7. Conclusion

This chapter has exposed consistent arguments backing the existence of a fundamental link between speech and music. It seems therefore quite reasonable to turn ourselves to music if we want to improve emotion recognition in speech.

3. Emotional Expression through Music

> *"Music is a moral law. It gives soul to the universe, wings to the mind, flight to the imagination, a charm to sadness, gaiety and life to everything. It is the essence of order and lends to all that is good and just and beautiful."* - Plato

In this chapter, we will take a look at the perception of music [46][86][79][18]. Some basic concepts of psychoacoustics like *critical bandwidth* or *differential threshold* shall be firstly introduced. We will then present musical concepts like *consonance*, *tension* or *modality* by looking at two-tone or three-tone phenomena. Finally, an experiment concerning tonality perception will be exposed which supports the idea of our auditory system being sensitive to distributional information in music.

3.1. Basics in psychoacoustics

3.1.1. Pitch perception

Pitch refers to the *perceived* fundamental frequency of a sound: it is a subjective psychophysical attribute. The total number of perceptible pitch steps in the range of human hearing is about 1400 distinct tones, whereas the total number of notes in the musical equal-tempered scale is 120 notes. Pitch is usually approximated by the fundamental frequency F0, although it is not exactly equivalent: it can happen, for example, that pitch is perceived even if the fundamental frequency is missing.

Pitch can be decomposed in two parts: pitch chroma and pitch height. Pitch chroma designates the actual note (ex: C, F, G...) whereas pitch height defines the octave in which the note occurs (ex: C can be a C1, C2, C3...). We call *pitch class* a set of all the pitches having the same chroma; that means, being a whole number of octaves apart. These pitch classes share similar "quality" or "color", which makes human pitch perception naturally periodic [100][1].

Pitch perception is a very complex phenomenon that can be influenced not only by the absolute frequency value, but also by the amplitude of the sound wave. For example, the pitch of a sinusoid increases with intensity when the sinusoid is above 3000 Hz. If on the other hand, the frequency of the sinusoid is below 2000 Hz, an increment in intensity is perceived as a drop in pitch.

In order to understand pitch perception, we shall focus on the inner structure of the ear. As it can be observed in Figure 3.1, there is a stiff structural element called basilar membrane inside the cochlea. The basilar membrane is the base for the sensory cells of hearing or "Stereocilia" (approximately 30.000 cells), and hence plays a crucial role in the

transfer of sound waves to the brain. It also fulfills the function of frequency dispersion for incoming sound waves. Depending on the input frequencies, different regions from the basilar membrane will resonate, activating only a small subset of sensory cells. Figure 3.1 shows that the distribution of frequencies is logarithmic rather than linear (which is why the function log appears constantly in speech or music processing).

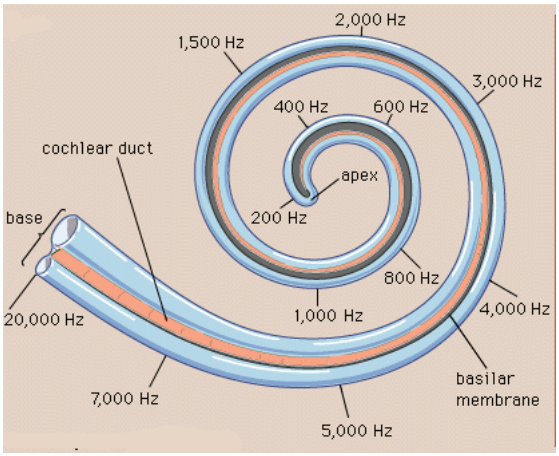


Figure 3.1.: Internal physiology of the ear

3.1.2. Critical bandwidth

An interesting question to ask ourselves is how our brain processes a superposition of complex tones. There are two kinds of superposition effects, depending on where they are processed in the listener's auditory system. If the processing is mechanical, occurring along the basilar membrane, we call them "first order superposition effects". "Second order" superposition effects are the result of neural processing and are more difficult to analyze. In this section, we will focus on the first order effects.

Let us consider two pure tones very close to each other, with equal amplitude and frequencies f_1 and $f_2 = f_1 + \triangle f$ respectively [73]. A famous psychoacoustic experiment consists in increasing the frequency difference $\triangle f$ and monitoring the perceived pitch; the different observed stages, illustrated in Figure 3.2, are the following:

- Unison: If $\triangle f = 0$, we hear one single tone f_1 whose amplitude depends on the phase difference between the two tones.
- Beating Effect: When we slightly increase the frequency f_2, an unpleasant sensation of *beating* will then appear. A single tone is still heard, but its frequency is $\frac{f_1+f_2}{2}$ and its amplitude is modulated by the frequency difference $\triangle f$. This phenomenon results from an overlapping of the two activated resonance regions in the basilar membrane.
- Roughness Effect: If the frequency difference $\triangle f$ is large enough, the beat

sensation disappears, but still a quite characteristic roughness or unpleasantness of the sound remains.

– TWO-TONES AREA: When $\triangle f$ surpasses the so-called *limit of frequency discrimination* $\triangle f_D$[1], we suddenly distinguish two separate tones of constant loudness corresponding to f_1 and f_2. Indeed, the two resonance regions on the basilar membrane are sufficiently separated from each other. This ability of the cochlea to distinguish a superposition of tones is called *frequency discrimination*.

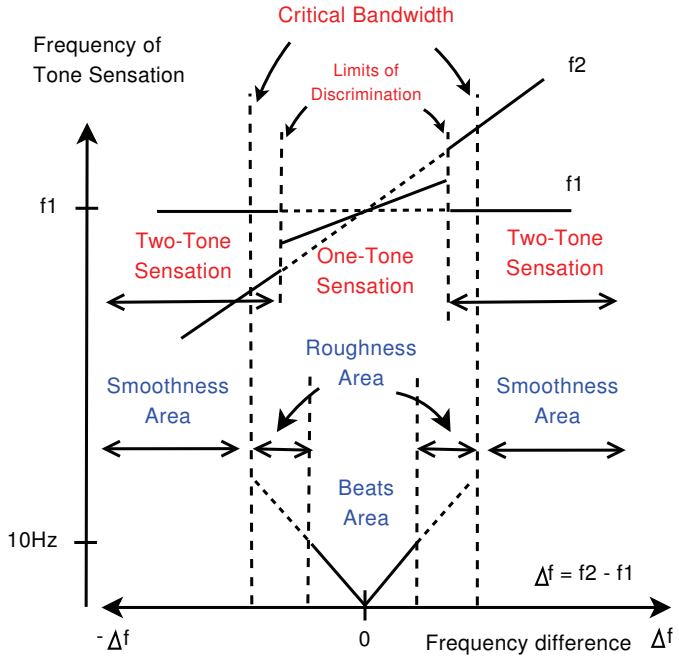


Figure 3.2.: Sketch of tone sensation caused by a two-tones superposition

It should be noted that the roughness and two-tone areas slightly overlap: the roughness sensation persists a little bit (specially in the low pitch range) even if the two tones can already be separated. Only after surpassing a yet larger frequency difference $\triangle f_{CB}$ called CRITICAL BANDWIDTH, the roughness sensation disappears, and both pure tones sound smooth and pleasing. Thus, the critical bandwidth corresponds to the frequency distance above which the roughness sensation disappears[2]. Our ear can be modeled as a filter bank logarithmically spaced. Critical Bandwidth would then refer to the bandwidth of each of those auditory filters. A critical band can also be understood as an information collection and integration unit on the basilar membrane, corresponding to a constant number of 1300 receptor cells.

[1]The limit of frequency discrimination should not to be confused with the Just Noticeable Threshold, which will be introduced in Section 5.4.1.

[2]This transition from "roughness" to "smoothness" is in reality more gradual: the critical band represents an approximate frequency separation.

3.1.3. Psychoacoustic scales

In order to study pitch perception, an appropriate scale should be selected according to the physiology of the ear [59]. Here is a list of the available scales:

Hertz Scale linear scale, close to physics. It is not appropriate for a perceptive analysis.

Bark Scale linear under 500 Hz, based on the concept of critical bandwidth. The Bark scale arises from covering the whole rank of frequencies with successive critical bands without overlapping. The Bark number from 1 to 24 corresponds to the 24 critical band of hearing. The formula to transform from Hertz to Barks is:

$$f_{Barks} = 13 \cdot \arctan\left(7.6 \cdot 10^{-4} \times f_{Hz}\right) + 3.5 \cdot \arctan\left(\left(f_{Hz}/7500\right)^2\right) \tag{3.1}$$

Equivalent Rectangular Bandwidth ERB between linear and logarithmic; this scale is similar to the Bark scale in the sense that it gives an approximation to the bandwidths of the filters in human hearing, but using the convenient simplification of modeling the filters as rectangular band-pass filters (unrealistic representation). The fundamental conversion transformation between ERB and Hertz is:

$$f_{ERB} = 11.17268 \times \log\left(1 + \frac{46.06538 \cdot f_{kHz}}{f_{kHz} + 14678.49}\right) \tag{3.2}$$

Mel Scale linear under 500 Hz, experimentally computed. This scale is commonly used for speech processing. It is a perceptual scale of pitches judged by listeners to be equal in distance from one another. Above about 500 Hz, larger and larger intervals are judged by listeners to produce equal pitch increments. The name "Mel" comes from the word melody to indicate its grounds on pitch comparisons.

$$f_{Mel} = 2595 \times \log_{10}\left(\frac{f_{Hz}}{700} + 1\right) \tag{3.3}$$

Semitone Scale logarithmic, based on music. This scale has been proved to be the most accurate to represent perceived pitch [59], and it will therefore be the chosen one for our experiments. The formula to transform the physical signal from Hertz to Semitone scale is the following:

$$f_{ST} = 69 + 12 \times \log_2\left(\frac{f_{Hz}}{440}\right) \tag{3.4}$$

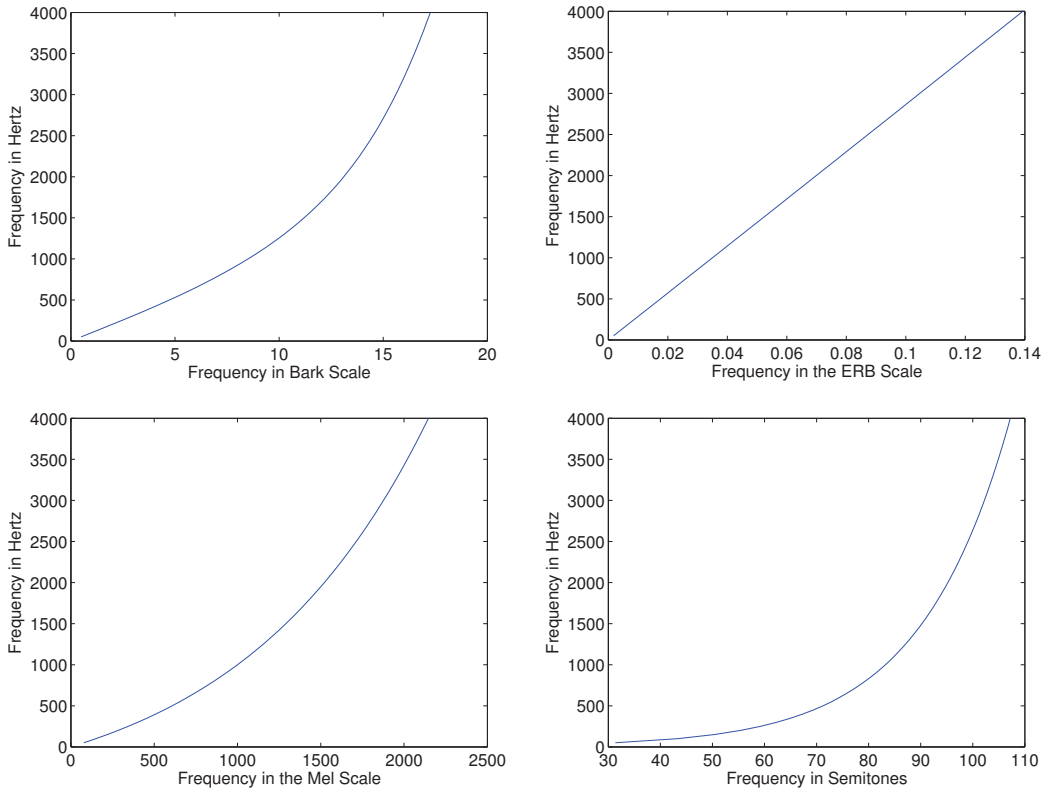


Figure 3.3.: Different psychoacoustic scales

3.2. Tonal consonance and critical bandwidth

> "*An unstable tone combination is a dissonance; its tension demands an onward motion to a stable chord. Thus dissonant chords are 'active'; traditionally they have been considered harsh and have expressed pain, grief, and conflict.*" - Roger Kamien, professor of musicology

Consonance in music refers to a sound resulting from an interval or chord that is pleasant, stable or resolved. Dissonance on the other hand is considered to be unstable, temporary or transitional (often associated to an unpleasantness when remaining in that chord). Pythagoras already noticed that intervals with ratios of 1:1, 1:2, 2:3 and 3:4 were more consonant than intervals with higher ratios. Imperfect consonances like 4:5, 3:5 or 5:8 appeared a little bit later, in the Middle Ages.

How can the relationship between consonance and frequency ratios be explained? According to the "Summation of interval effects" theory, consonance is actually due to the absence of rapid beats between harmonics of the component tones. Beats and roughness appear for small frequency differences between two tones, which is why consonance is strongly related to the concept of critical bandwidth, defined previously in Section 3.1.2.

The idea that consonance depends on the number of coinciding harmonics is well-known; yet, recent studies have proved that frequency distance is more important than frequency ratio in order to understand consonance. Let us consider two pure sinusoid tones (no harmonics are involved yet). Dissonance will appear if these two tones are close enough, inside the critical bandwidth. Figure 3.4 plots the interval consonance between two pure tones versus the relative distance expressed in percentage of the critical bandwidth. Maximal tonal dissonance is produced by interval distances at 25% of the critical bandwidth, whereas maximal consonance is reached for interval widths above 100% of the critical bandwidth. The fact that distances are expressed in percentage of the critical bandwidth makes the consonance degree nearly independent of the absolute frequency value.

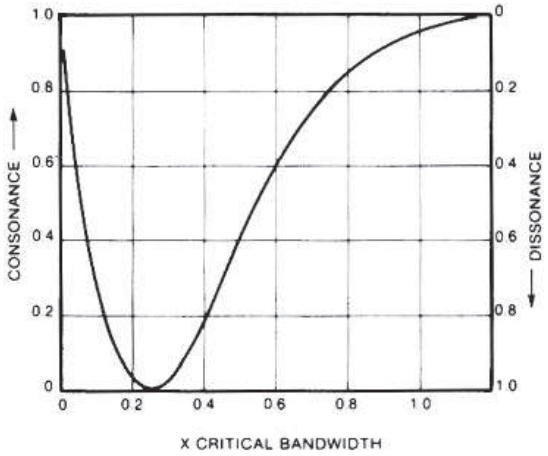


Figure 3.4.: Dissonance versus critical bandwidth between two *pure* tones

For complex tones, we have additionally several partials or harmonics for each tone. In order to compute the dissonance, we should apply the idea of Figure 3.4 to each pair of harmonics. The total dissonance will therefore be the sum of individual dissonances for each pair of adjacent partials, resulting in the representation of Figure 3.5.

It can be observed that frequency distances corresponding to simple frequency ratios obtain high consonance ratings. The simpler the frequency ratio is, the sharper the peak is (the width of the peaks explains the tolerance of our ear to impure intervals). Moreover, the usual ranking order of consonant intervals agrees quite well with the relative heights of the peaks. To sum up, the relationship between consonance and small frequency ratios can be explained by the actual distance between harmonics. Every time that the frequency difference between two partials is under the critical bandwidth, unpleasant beatings appear and a dissonance factor will then be added.

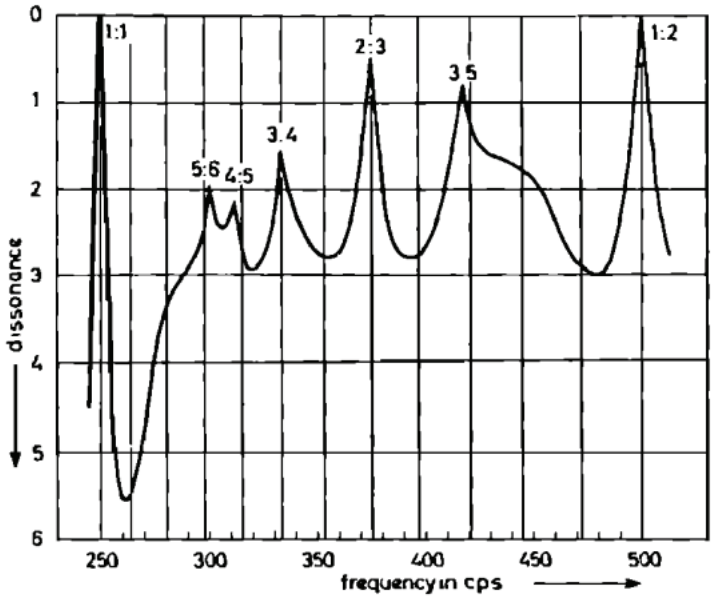


Figure 3.5.: Dissonance versus distance in frequency between two *complex* tones

3.3. Triad perception theory

In music, harmony is usually studied in terms of three-tone chords called *triads*. Figure 3.6 illustrates the four most important triads. It is well-known that each three-tone chord has different attributes and degree of consonance. If we order the triads from the most to the less consonant one, the obtained sequence is: major, minor, diminished and augmented[3]. Moreover, major and minor chords have always been associated with an affective valence, either positive or negative: these chords are at the basis for the diatonic and pentatonic musical scales worldwide [83].

In the previous section, dissonance has been explained as the result of individual dissonances among tones and their upper harmonics. Yet, the perceptual ordering of the triads or the fact that some chords have a positive or negative connotation cannot be explained solely with an interval-based model. In [10], Cook and Fujisawa introduce a new model accounting for three-note effects. According to this new theory [9][12][10], triads can be classified based on the concepts of tension and consonance. *Consonance* has been defined in Section 3.2, and *Tension* refers to the perceived need for relaxation or release created by a listener's expectations. Figure 3.7 presents a classification of the triads into three distinct categories: sonorous chords (containing unequal, consonant intervals), tense chords (containing "intervallic equidistance") and dissonant chords (containing one or more dissonant intervals).

Now, let us compare the size of the upper and inferior intervals. Tension appears whenever these two intervals have an equivalent size, and it can only be resolved if we

[3]Inversion of triads can change their perception.

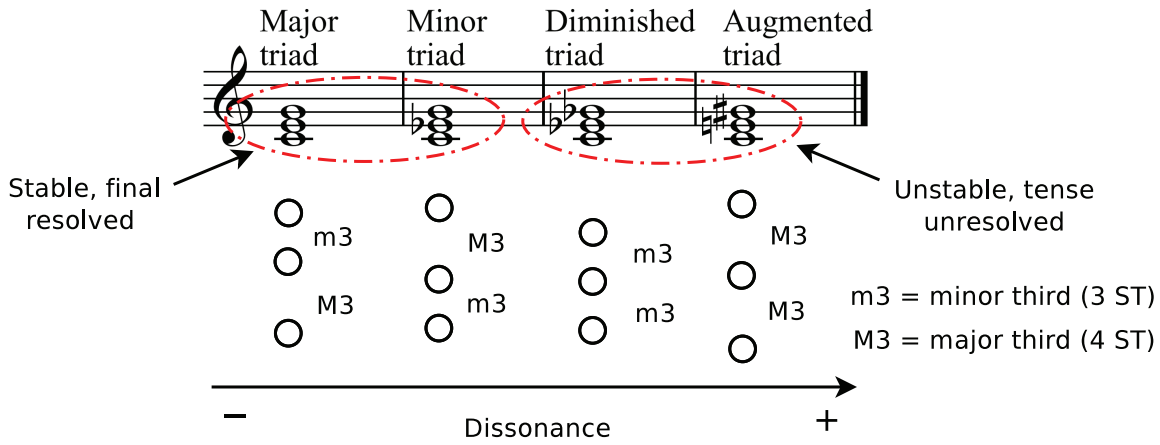


Figure 3.6.: Types of triads

change one of the intervals in order to have unequal sizes. Intuitively speaking, the middle tone seems to be "trapped in the middle" whenever the triad is equally spaced. The stability of the chords is hence influenced by both two-tone effects (consonance versus dissonance) and three-tone effects (sonority versus tension).

In the following, let us consider three tones with frequencies f_1, f_2 and f_3.The instability I of any three-tone chord can be expressed as

$$I = D + \delta \cdot T \tag{3.5}$$

where D stands for the overall dissonance, T refers to the overall tension between the three notes and δ is a constant term experimentally deduced. The overall dissonance D will be expressed as the summatory of individual dissonance terms d between harmonics, like this:

$$D = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} d\left(x_{ij}, v_{ij}\right) \tag{3.6}$$

where $x_{ij} = \log\left(f_j / f_i\right)$ refers to the interval of frequency between harmonics i and j, and v_{ij} is the product of the relative amplitudes of the two tones. The individual dissonance can be modeled with the following equation, according to the experimental graph of Figure 3.4 on page 27:

$$dissonance\ d = v \cdot \beta_3 \left\lfloor \exp\left(-\beta_1 x^{\gamma}\right) - \exp\left(-\beta_2 x^{\gamma}\right) \right\rfloor \tag{3.7}$$

where $x = \log\left(f_2 / f_1\right)$, v refers again to the product of the relative amplitudes of the two tones and γ, β_1, β_2 and β_3 are experimentally deduced parameters.

As far as the overall tension T is concerned, it also corresponds to a summatory of individual tension factors between harmonics of the three tones, like this:

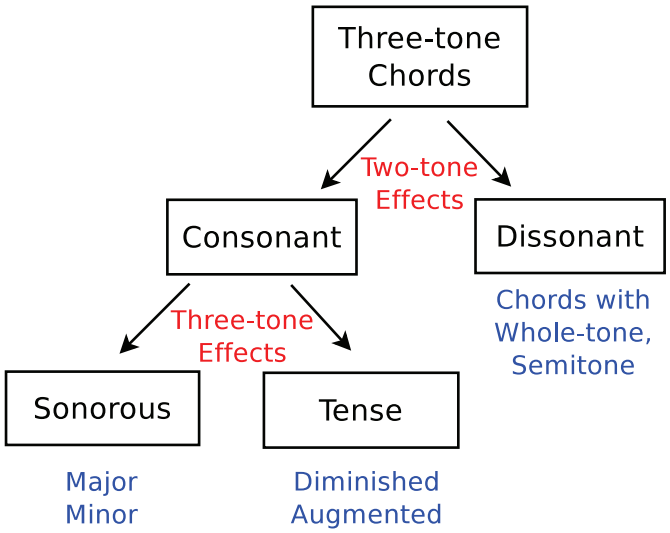


Figure 3.7.: Classification of triad chords based on tension and consonance

$$T = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} t\left(x_{ij},\, y_{jk},\, v_{ijk}\right) \tag{3.8}$$

The individual tension t between three *pure* tones can be modeled as:

$$tension\ t = v \cdot \exp\left[-\left(\frac{y-x}{\alpha}\right)^2\right] \tag{3.9}$$

where $x = \log\left(f_2/f_1\right)$ and $y = \log\left(f_3/f_2\right)$. v is the product of the relative amplitudes of the three partials. The parameter α is experimentally derived.

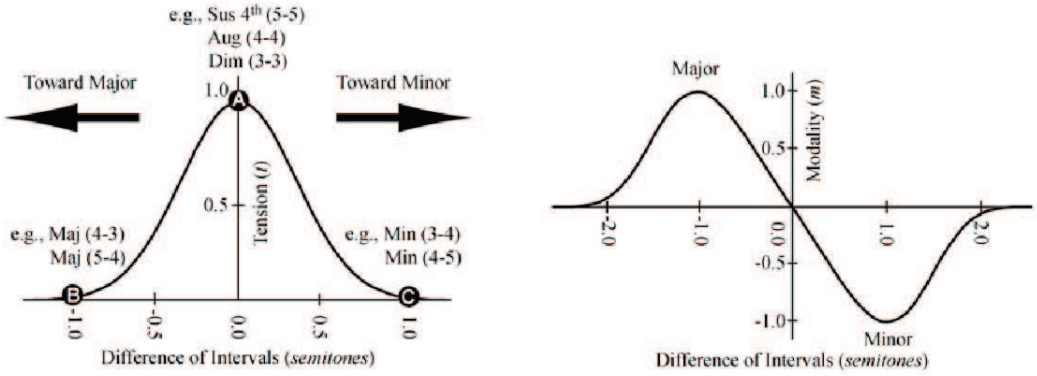
Last but no least, an empirical model to explain the valence of major and minor chords will be exposed. Indeed, major and minor chords represent the only two possible resolutions of chordal tension: the pitch of the middle tone can either go up or down. Here again, the overall modality will be the sum of individual modality factors between the harmonics, and it can be given as:

$$M = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} m\left(x_{ij},\, y_{jk},\, v_{ijk}\right) \tag{3.10}$$

$$modality\ m = -v \cdot \left[\frac{2\left(y-x\right)}{\varepsilon}\right] \exp\left\{-\left[\frac{-\left(y-x\right)^4}{4}\right]\right\} \tag{3.11}$$

where v again accounts for the relative contribution of the three partials, x and y are the lower and upper intervals, respectively, and the parameter ϵ is set so that the modality value will be +1 for the major chord and -1 for the minor chord.

Figure 3.8 presents a schematic modelization of the concepts Tension and Modality. In both sketchs, the abscissas correspond to the frequency difference $\triangle f = \text{interval}_{\text{upper}} - \text{interval}_{\text{inferior}}$ in semitones between the upper and inferior intervals of a triad. The tension of the three-tone chord will be maximum for equally spaced intervals ($\triangle f = 0$); it then decreases for an increasing $\triangle f$: depending on with interval will be bigger, we go toward the major or minor modes. In the case of modality, the curve behavior is circularly symmetric: when $\triangle f = -1$, the inferior interval is 1 semitone bigger and we obtain a major chord. If however $\triangle f = 1$, the upper interval is the biggest and we are dealing with a minor chord.



(a) Tension model

(b) Modality model

Figure 3.8.: Tension and modality as a three-tone perception

3.4. Perception of hierarchical structure of music

In music, we call *Tonality* (or musical key) a hierarchical ordering of the pitches of the chromatic scale such that these notes are perceived in relation to one central and stable pitch called the tonic. This feeling of hierarchical ordering is essential for the transmitted musical message: the same sequence of notes can actually induce completely different sensations depending on the tonality used [8][92].

Given a certain tonality, it is generally acknowledged that some notes are more representative than others. This order of importance can be represented as a distributional function called Standardized Key Profile (SKP). Figure 3.9 shows an example for the musical key major C. Stable pitches are generally played most frequently and have greater tonal duration. Apparently, listeners are sensitive to distributional information in music, which allows algorithms for tonality identification to be based on SKP-matching. In [91], the properties underlying our sensitivity to distributional information were investigated. The final observations of this study were the following:

– Organization and Differentiation: Tonality perception is indeed affected by the pitch hierarchical structure, but there is also a minimum differentiation

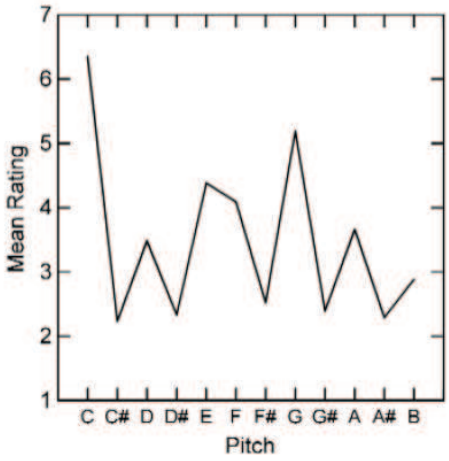


Figure 3.9.: Standardized Key Profile for C dur

required. If we multiply the SKP by an attenuation factor, the distributional function might be too flat to be perceived as a distinct tonality.

- The hierarchical organization of pitches is build on the basis of NOTE DURATION (rather than frequency of occurrence). The selective attention to an element is proportional to its duration.
- MULTIPLE HIERARCHIC LEVELS are required for an adequate tonal perception (if we differentiate only the tonic from the non-tonic pitches, the tonality sensation is not fully created).
- RANDOM ORDERINGS of notes can invoke the same tonality in listeners, under the condition that they follow the appropriate tonal distribution.

Finally, the fact that tonality perception is related to distributional information in music may reflect a very general principle of our auditory system. Not only it is important to have a minimum level of tonal magnitude, it is also required to have multiple hierarchic levels of pitch organization.

3.5. Conclusion

We were wondering in a first place how music is able to induce emotions at all. The answer of philosophers like Stephen Davies or Jerrold Levinson lies in the concept of *appearance emotionalism*: some musical features are indeed similar to characteristics of human affects (for example, a sad person will probably move more slowly than usual; similarly, music that induces a sad feeling tends to have a slower beat).

But is it possible to identify certain acoustical characteristics associated to particular emotional states? Generally speaking, there is no such universal reaction patterns

to music: each individual becomes influenced by his own memory, personality and environment). There are nevertheless some acoustic properties of musical signals that can define uniquely basic emotions, transcending individuality and acculturation. Such acoustic properties (or at least a subset of them) have been exposed in this chapter, and they partially explain how musical emotions are induced:

1. Consonance and Dissonance (Interval Consonance Theory in Section 3.1.2).
2. Tension and Modality (Triad Perception Theory in Section 3.3)

   Consonance, dissonance, tension and modality have a strong link to the valence dimension in the three-dimensional emotional space (see Figure 1.3).
3. sensitivity to distributional information constitute an important property of our auditory system. It conditions our perception of the exposed musical concepts (Section 3.4).

Other interesting theories concerning emotional expression are based on unexpected harmonic changes. Emotions would result from the violation of musical expectations of the listener. Interesting models of tension accumulation across musical sentences and musical syntax can be found in [47].

4. Emotion Recognition from Speech Signals

"*The essential difference between emotion and reason is that emotion leads to action while reason leads to conclusions.*" - Donald Calne, neurologist

In this section, the problem of emotion recognition from speech signals will finally be approached [15][97]. General concepts about speech will first of all be introduced; we will then describe the general problem of pattern recognition from which emotion recognition is a concrete example. The most common features (called basic feature set) will then be presented. This section concludes with a theoretical overview of methods for feature evaluation.

4.1. Human speech characteristics

Source-filter model Speech production has often been modeled as a linear convolution between source and filter (see Figure 4.1). Speech signals result from air pressure variations produced by the vocal system. The lungs provide the power (air) to the system, and the vocal folds located in the larynx generate the principal sound of speech, with fundamental frequency F0. The filtering step is finally performed by the vocal tract, which attenuates or enhances certain frequencies through resonance effects. The final spectrum of a speech signal can hence be treated as the product of an excitation spectrum and a vocal tract spectrum.

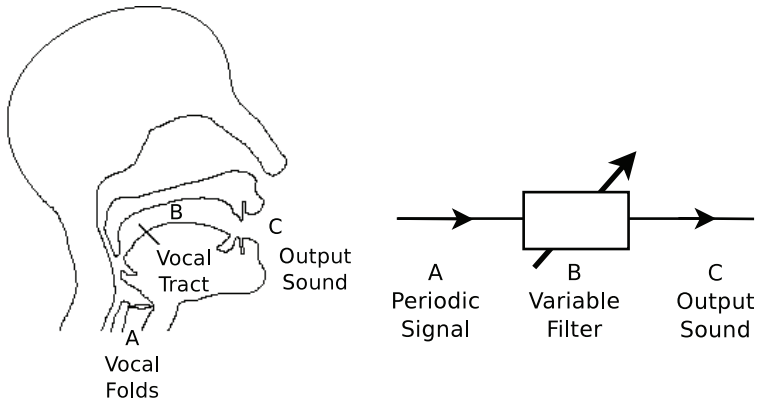


Figure 4.1.: Source-filter model

Formant versus harmonic We call *formant* a concentration of acoustic energy around a particular frequency in the speech wave. Formants and harmonics should not be confused: whereas harmonics are component frequencies of the signal that are integer multiples of the fundamental frequency F0, formants appear from modulations of the vocal tract. During vowel sounds, most of the energy is concentrated in the first three or four formants. These concepts have been illustrated in Figure 4.2.

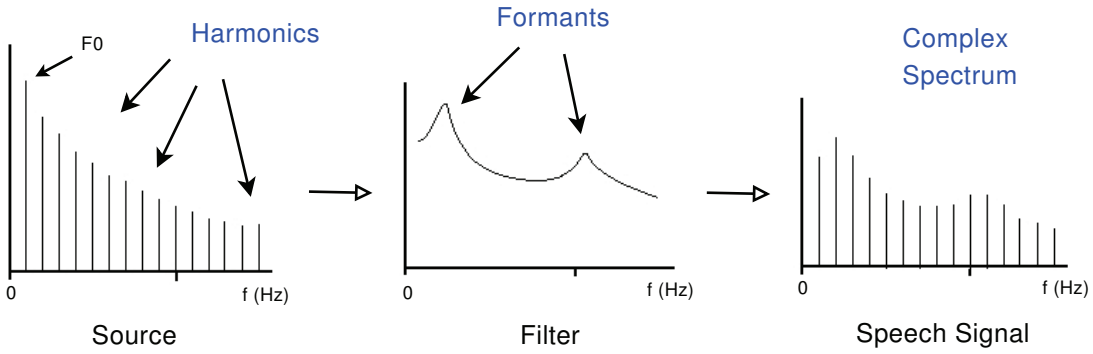


Figure 4.2.: Illustration of formants

Voiced versus unvoiced speech In a speech utterance, we have spoken and silent parts. Among the spoken segments, these can be either voiced or unvoiced. Voiced speech tends to have a strong periodicity, and it roughly corresponds to vowels. It is caused by periodic pulses of air generated by the vibrating vocal folds. About two-thirds of speech is voiced and it constitutes the most important part for intelligibility. On the other hand unvoiced speech refers to a noisy and non-periodic signal caused by air passing through a narrow constriction of the vocal tract as when consonants are spoken.

	Voiced	Unvoiced
ZCR	low	high
Energy	high	low
Energy Concentration	$<$ 1kHz	$>$ 1 kHz
Wave Shape	periodic	noisy
Autocorrelation	high	low
Nature of Phoneme	vowel	consonant

Table 4.1.: Principal characteristics of voiced/unvoiced segments

Table 4.1 presents a comparison between voiced and unvoiced characteristics. ZCR

refers to Zero Crossing Rate and will be defined in Section 4.4. It can be seen that ZCR is low for voiced segments and high for unvoiced ones, whereas energy is high for voiced blocks and low for unvoiced ones. It is also interesting to note that energy for voiced parts concentrates in the frequencies below 1kHz, while energy for unvoiced parts concentrates in the higher bands of the spectrum [30][36].

4.2. Pattern recognition

Pattern recognition is a field in statistics whose aim is to recognize sub-patterns within a big amount of data. It entails extracting the most relevant information associated to those patterns, which allows the classification of new data. According to the type of learning procedure, pattern recognition problems can be classified in four different families, namely supervised, reinforced, semi-supervised and unsupervised learning.

In supervised learning, we work with labeled training data, which is very desirable. Unsupervised learning on the other hand assumes no a priori information about the classes: the objective is to find inherent patterns or clusters which would help for classification of new data. Since labeling data is a very expensive task, semi-supervised learning arises as a combination of both. The training data is hence partially labeled (typically a small set of labeled data together with a large amount of unlabeled data).

Finally, reinforced learning is concerned with how an agent ought to take actions in an environment so as to maximize some notion of cumulative reward. In this case a small feedback instead of correct input/output pairs is presented. In this thesis, we will be restricted to situations of supervised learning, with audio files labeled with emotional classes. Figure 4.3 shows the needed steps to solve a general supervised pattern recognition problem.

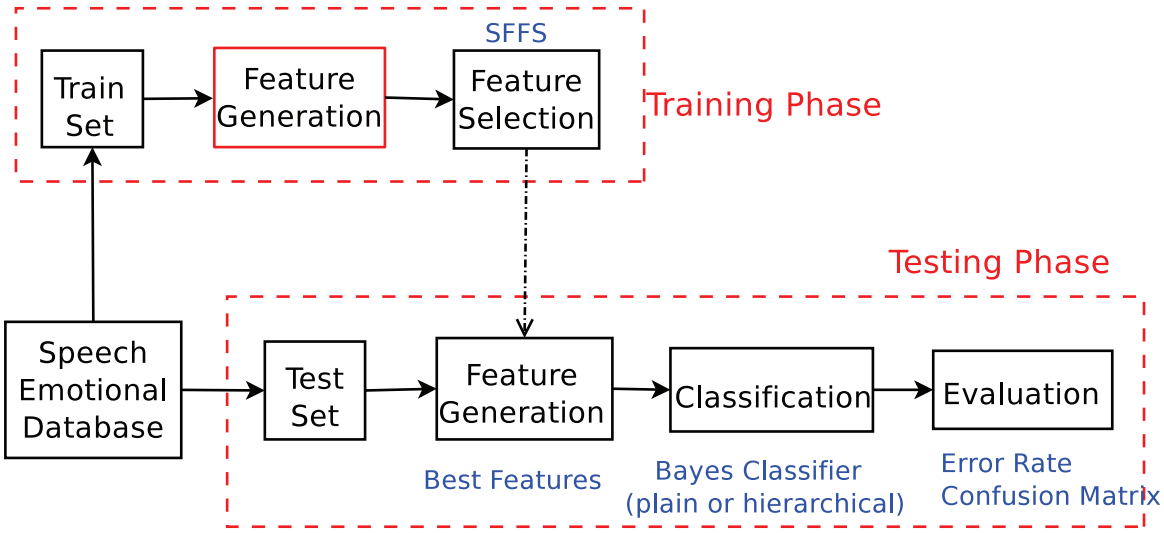


Figure 4.3.: Architecture for pattern recognition

Split Train/Test If we want our system to be speaker independent, a *one-speaker-out evaluation* shall be performed. That means that the splitting has to be done for each user, taking all the others for training and the selected one for testing. A second possibility is to conduct an *m-fold cross-validation,* which consists in splitting randomly the data in m sets and running the experiment for each set, in order to get statistically stable results. In this case, the system is speaker-dependent.

Feature Generation This step is a data mining procedure to reduce the amount of data and to find the most meaningful information. This is the main focus of the thesis. A detailed description of the extracted features will be given in the next section.

Feature Selection Among the complete set of features, it is imperatively required to make a selection of the best ones, otherwise overfitting will appear. In our case, the Sequential Floating Forward Selection (SFFS) algorithm will be used. It is an iterative method which tries to find a subset of features close to the optimal one. At each iteration, a new feature is added to the previous subset (forward step). Afterwards, the least significant features are removed as long as we obtain better recognition rate than in previous subsets with the same number of features (backward step).

Classification Using the selected features, the classifier predicts the best emotional label for each sample. So far, there exist two general tendencies: either we generate a big amount of features and apply a complex classifier like Support Vector Machines, or we use a moderate number of meaningful features together with a rather simple classifier like Bayesian-Gaussian Mixture Model.

In our case, we have decided to use plain and hierarchical Bayes classifiers. The hierarchical classifier consists of 5 binary Bayes classifiers that distinguish between the three dimensions (activation, potency and valence) individually. The configuration that has been chosen is represented in Figure 4.4.

Evaluation In order to evaluate the performance of our system, we will compute the error rate, weighted error rate and confusion matrix, comparing the predicted and actual emotion class labels.

Error Rate It simply consists in counting how many labels were wrongly assigned. Its opposite is called *Accuracy* or *Recognition Rate*. When a data set is unbalanced (that is, the number of samples vary greatly between classes) the error rate of a classifier is not representative of the true performance of the classifier. For this, we use other measures like weighted error rate or confusion matrix.

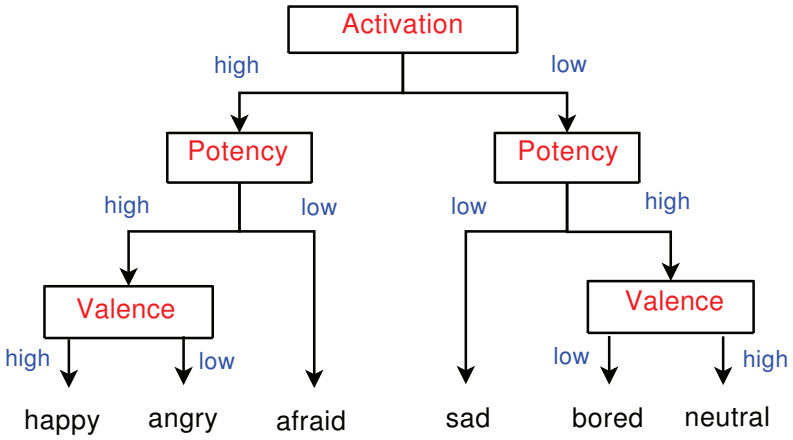


Figure 4.4.: Hierarchical Bayes classifier

Weighted Error Rate It takes into account how many errors there are, but this time errors for each class are weighted according to the size of each class. If for example we have a classifier which would always predict the same biggest category, all the samples of that category will be counted as a hit, but the classifier is clearly undesirable. This measure penalizes such classifiers.

Confusion Matrix This is the most detailed measure. It captures the local errors by representing the predicted classes in rows and the actual classes in columns.

As a remark, an additional block called *Feature Transformation*[1] has been omitted in the scheme since it was absent in our experiments. The objective of this block is to transform the features in order to get a better catch of their essence (detect cross-correlation or redundant information). This block constitutes an interesting path for future research (see Section 7.2.2).

4.3. Feature generation

Feature generation is a three-step process: the first step is a pre-processing stage in which the speech utterance has to be normalized and eventually noise-filtered or smoothed. The second and third step correspond to the extraction of *local and global features*, as can be seen in Figure 4.5. Features can indeed be different in nature, depending on whether they capture local or global information. A speech utterance is typically segmented into small blocks in which local computations are performed (ex: energy, pitch of the window).

Global features refer to all the computations performed with local features; they use information from the whole utterance, whereas local features are calculated for each block independently[2]. The easiest ones are statistical overall values such as mean, maximum,

[1]This block is analog to the Feature Selection block. It could be performed before, after or instead of this one.

[2]In our implementation, we have one value of the local feature per block.

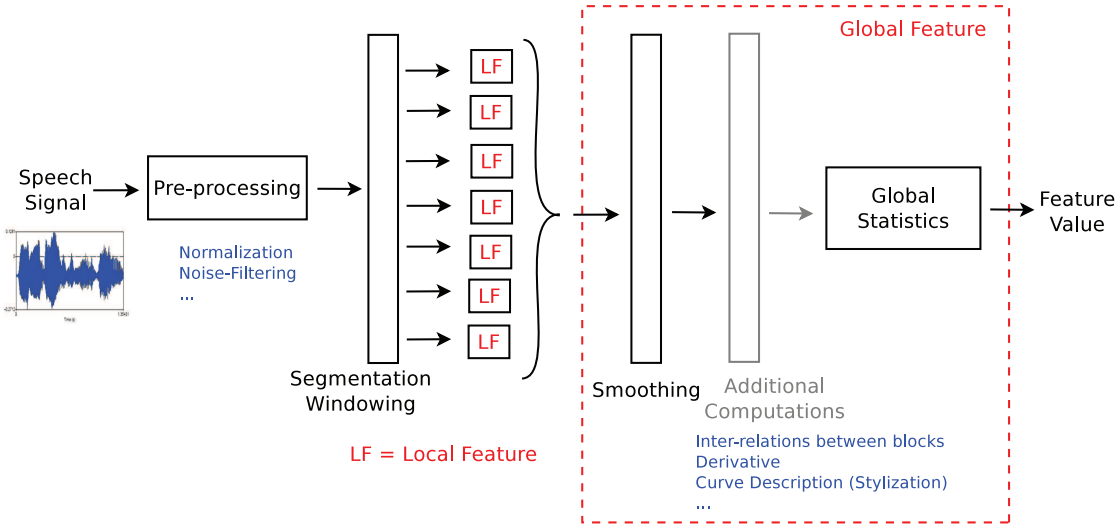


Figure 4.5.: Procedure for feature extraction

minimum, median, variance, range or interquartile range. More complex features include some in-between operations like the following:

- curve simplification in linear segments like slopes or constant levels (ex: stylization process of the pitch, macroscopic variations of the energy, etc...)
- combination of multiple local features (for example, mean energy of the voiced segments)
- histogram derivation (for musical features)

4.4. Traditional features

In the following, a detailed description of the most common features in the field of speech emotion recognition will be provided [3][15][84][74][27]. The features concerning duration, energy and pitch are directly correlated with tempo, loudness and intonation respectively.

Duration: These features provide temporal properties about voiced and unvoiced segments. They operate directly on the temporal signal (ex: mean duration of the voiced segments).

Mel-Frequency Cepstrum Coefficients (MFCCs): These coefficients result from a transformation to a cepstrum space, in order to capture information of the time-varying spectral envelope. A cepstrum can be obtained by applying a Fourier Transform on the $\log(f)$ plot, in order to separate in the frequency domain the slowly varying spectral envelope from the more rapidly varying spectral fine structure (separation

of the source and filter spectrum). We are interested in the filter spectrum which varies more slowly and carries valuable prosodic information.

$$power\,Cepstrum = \left| FT\left\{ \log\left(\left| FT\left\{ x\left(n\right)\right\} \right|^2 \right) \right\} \right|^2 \tag{4.1}$$

Energy: The energy of a signal x in a certain window of N samples is given by:

$$En = \sum_{n=1}^{N} x\left(n\right) \cdot x^{\star}\left(n\right) \tag{4.2}$$

Zero Crossing Rate (ZCR): The Zero Crossing Rate counts how many times the speech signal changes its sign:

$$ZCR = \frac{1}{2} \cdot \sum_{n=1}^{N} \left| \mathrm{sgn}\left(x_n\right) - \mathrm{sgn}\left(x_{n+1}\right) \right| \tag{4.3}$$

Pitch: The pitch or fundamental frequency F0 along time can be estimated in multiple ways. In our case, we have used two different algorithms based on the short-time autocorrelation, one very immediate but not very exact and another more accurate but also more complex (Robust Algorithm for Pitch Tracking), which will be explained in Section 4.5.

Formants: These features capture spectral information about the position and shape of the formants. It has traditionally been used for Automatic Speech Recognition (ASR), but it also brings emotional information related to articulation. Algorithms for pitch and formant tracking are very similar.

Voice Quality Parameters: Not so classical, these features are closely related to phonation and are based on an acoustical model of the vocal folds. They are calculated by first inverse-filtering the speech signal. The influence of the vocal tract is hence compensated to a certain extent and an estimate of the vocal folds vibration is obtained. Afterwards, several spectral gradients of the glottal signal in the frequency domain can be calculated.

All these traditional acoustic parameters still present some difficulties to distinguish happiness from anger or sadness from boredom. Our aim therefore will be to improve results by focusing on the melody and harmony of speech (see Chapter 5). This approach is possible under the assumption that our brain integrates temporally what we hear in a short period of time of at least 3 or 4 seconds. It is thus possible to analyze pitch intervals or triads as if they occurred simultaneously, even though speech is by nature sequential.

4.5. Pitch estimation algorithm

> "Although emotion can be expressed at all levels, from the semantic to the prosodic, the most immediate expression of emotion in speech is through *pitch movement*" - Levelt, researcher in Psycholinguistics

Pitch plays an essential role in the transmission of vocal emotions [61][62][70][6]. Any algorithm for pitch estimation can be decomposed in two principal steps: the first one finds potential F0 candidates for each window and the second one selects the best candidates and eventually refines the estimation [31][7]. Most pitch estimation algorithms can be grouped in three categories, as listed in Table 4.2 on page 45.

All these algorithms can moreover be improved in two different ways. The first one consists in adding a *modelization of the auditory system*. An example is the use of a different spectral transform instead of a basic Fourier transform, which would be more faithful to the perceptual response. A second improvement is to apply *tracking*, that is, taking into account the surrounding values for F0 at each moment. Indeed F0 varies continuously and slowly, and therefore actual values should be close enough to immediately preceding or following ones (ex: a Hidden Markov Model can be built).

Robust Algorithm for Pitch Tracking (RAPT)

This algorithm is based on the computation of the autocorrelation. Algorithm 4.1 and Figure 4.6 give respectively a description and graphical representation of the process. A detailed description of the algorithm can be found in [43].

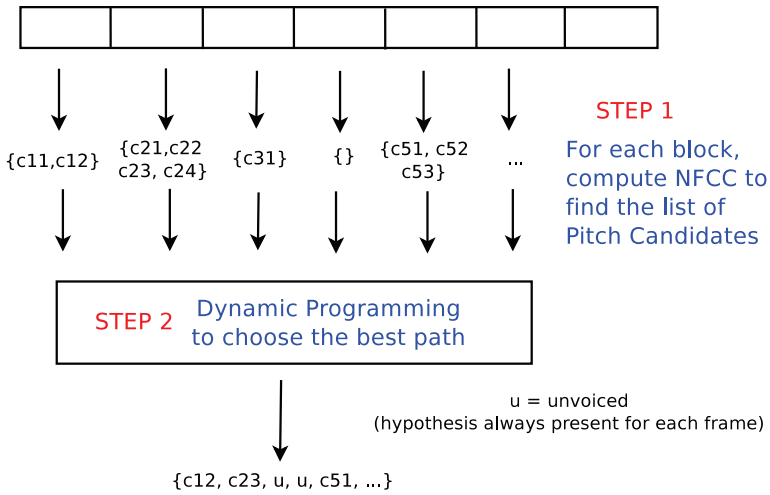


Figure 4.6.: Implementation of the RAPT Algorithm

The first stage computes the Normalized Cross-Correlation Function ($NCCF$) for each block in order to find possible candidate estimates for F0. The $NCCF$ is given by

$$NCCF_{i,k} = \frac{\sum_{n=m}^{m+N-1} x(n)\, x(n+k)}{\sqrt{e_m e_{m+k}}} \tag{4.4}$$

$$e_m = \sum_{n=m}^{m+N-1} x\,(n)^2 \tag{4.5}$$

where $x\,(n)$ is the windowed signal, m refers to the index of the frame i and k is the lag for the autocorrelation. Normalizing the cross-correlation compensates for possible energy differences along the whole signal.

The second step applies dynamic programming between all frames in order to choose the best path for F0. The best candidates for each frame are selected based on a combination of punctual and contextual information. Local and transition costs are calculated according to psychoacoustic properties of our ear. The path with minimum cost will be chosen at the end.

Algorithm 4.1 Robust Algorithm for Pitch Tracking (RAPT)

1. $NCCF$ to generate candidate F0 estimates
   a) Provide two versions of the speech signal; one at the original rate, another at a significantly reduced rate.
   b) Compute the $NCCF$ per block on the low-sample rate signal version. Record the location of local maxima.
   c) Compute the $NCCF$ per block on the high-sample rate version only in the vicinity of the maxima. Search again for refined maxima (improvement of peak location and amplitude estimates).

   $\Longrightarrow$ Each maximum from the high-resolution $NCCF$ generates a candidate F0 for that block. Also the hypothesis of unvoiced frame is always considered.
2. Dynamic programming to select the best candidates across all frames
   a) For each block i, calculate iteratively the total local cost as $C_{ij_1} = C_{(i-1)j_2} + transitionCost_{j_1 j_2}$, where j_1 and j_2 refer to F0 candidates. Proceed until having the total cost of each path.
   b) Choose the path that minimizes the total cost and take its nodes as final pitch estimation.

4.6. Methods for feature evaluation

In order to evaluate how good our generated features are, we have two possible strategies: feature ranking or feature dimensionality reduction. Feature ranking consists in assigning values to our features that weigh up their relevance for our particular problem. We can then give the K-top features (the ones with highest scores) to our classifier.

On the other hand, feature dimensionality reduction consists in reducing the feature vector $\underline{x} \in \mathbb{R}^N$ to a new feature vector $\widetilde{\underline{x}} \in \mathbb{R}^K$ with $K < N$ under some optimization

criteria. In our simulations we have chosen this strategy (usage of the Sequential Floating Forward Selection SFFS algorithm), which gives the best classification results in despite of an increase in computational cost. Here we present two examples of feature ranking techniques and compare them with the SFFS.

Univariate feature ranking

By univariate, we mean that we will consider each feature individually. One of the most famous examples is the Fisher coefficient. In this case, each class is represented by its centroid c_j. The Fisher coefficient can then be written as

$$fisher = \frac{between\,class\,scattering}{within\,class\,scattering} \tag{4.6}$$

$$= \frac{\sum_{j=1}^{c}\sum_{x_i \in w_j}(x_i - c_j)^2}{\sum_{j=1}^{c} N_j\,(c_j - c)^2} \tag{4.7}$$

where

$$c = \frac{1}{N}\sum_{n=1}^{N} x_i \tag{4.8}$$

If the Fisher coefficient is large, it means that we will have a good separation of classes: the feature is a good one. If however this coefficient is small, the classes are not properly separated and the feature alone is not very informative. The problem of this coefficient is that each class is represented by a single centroid: if a class is multimodal (more than a single cluster), a single centroid is not representative.

Multivariate feature ranking

Univariate feature ranking methods present the important disadvantage of redundancy between features. Since we are considering each feature individually, we might get high scores but not necessary new information. We should instead take features that complement each other: features that are individually poor may show a good performance together.

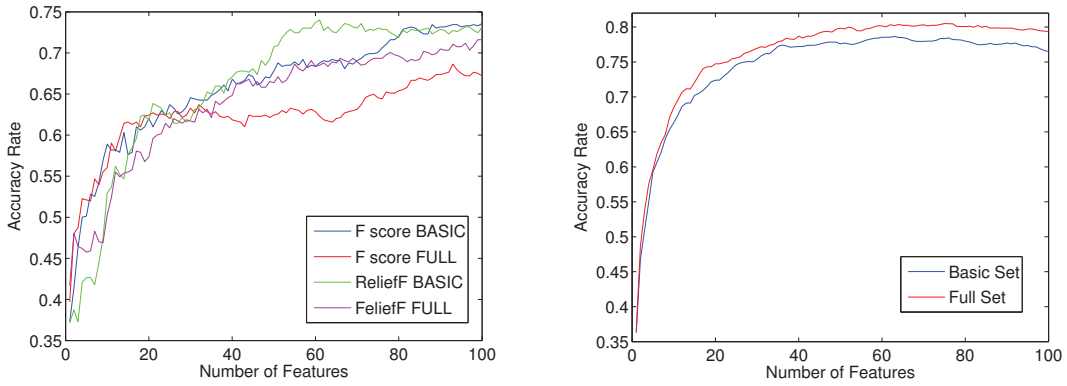
Hence, multivariate feature ranking methods consider the relationship between different features, removing redundancy between classes. An example would be the ReliefF algorithm: it selects data points randomly and calculates the nearest hit and nearest miss values depending on their neighborhood. A detailed description can be found in [90].

Comparison of methods

Figure 4.7 compares the performance of the different exposed methods for both the basic and full (basic and musical) sets of features. Fisher coefficient (F score in the plot) is the worst measure: this was expected, since we are considering each feature

individually, thus removing possible inter-feature information and keeping possible inter-feature redundancy.

It is interesting to notice that in the case of feature ranking, the basic set of features gives better results than the full one: the musical features seem to have redundant information in respect to the basic ones. The results of the SFFS algorithm are the most accurate. This is why we always use the SFFS method in our simulations.



(a) Feature ranking

(b) Feature dimensionality reduction

Figure 4.7.: Comparison of feature evaluation methods

4.7. Conclusion

In this chapter, we have presented the problem of emotion recognition from speech signals. The general pattern recognition approach and the feature generation step have been explained. We have then described the traditionally used features, our pitch extraction algorithm and finally a short review of methods for feature evaluation. The next chapter will present the core of this thesis, namely the musical features.

Time-Domain Methods
– Time-Event Rate Detection: Detect events like ZCR, peak rate, slope rate. Easy to understand, simple but limited accuracy. – Autocorrelation: RAPT, YIN Algorithms. – Phase Space: Plot the signal in phase space and observe its frequency. Periodic signals draw closed cycles in phase space, and F0 is related to the speed of such cycles. The problem appears when the cycle has re-tracing or crossing parts.
Frequency-Domain Methods
– Component Frequency Ratios: For each pair of partials, find the smallest "harmonic numbers"; robust method, it works also with missing F0, missing partials or inharmonic partials. – Filter-based Methods: Filter the signal, either with an optimum Comb Filter or a tunable IIR Filter; robust but computationally expensive. – Cepstrum Analysis: Compute the cepstrum of the signal. Under the assumption that the signal has regularly-spaced frequency partials, good for speech processing. – Muti-resolution Methods: Apply different time window sizes to calculate the spectrum. Discrete Wavelet Transform is particularly fast.
Multi-Resolution Methods
– Neural Networks – Maximum Likelihood Estimators These methods model the human auditory system. The disadvantage is their black-box behavior; there are no clear explanations.

Table 4.2.: Pitch estimation algorithms

5. Musical Features

> "*Music expresses feeling and thought, without language; it was below and before speech, and it is above and beyond all words.*" - R. Ingersoll, orator

This chapter describes the musical features implemented in this thesis, as well as possible ideas and hypotheses for future research paths. The first part introduces tonal distribution features which are based on the histogram of the circular pitch autocorrelation [53], illustrated in Figure 5.1. In this case, a previous normalization of the pitch is performed by subtracting the mean pitch of the signal, as well as a conversion from Hertz to Semitone scale. Secondly, several features used for music emotion recognition have been implemented for speech signals, namely intensity, timbre and rhythm features. Finally, we have generated new features based on perception theory. These features will be generated from a new perceptual pitch signal, resulting from the stylization of the physical F0 signal.

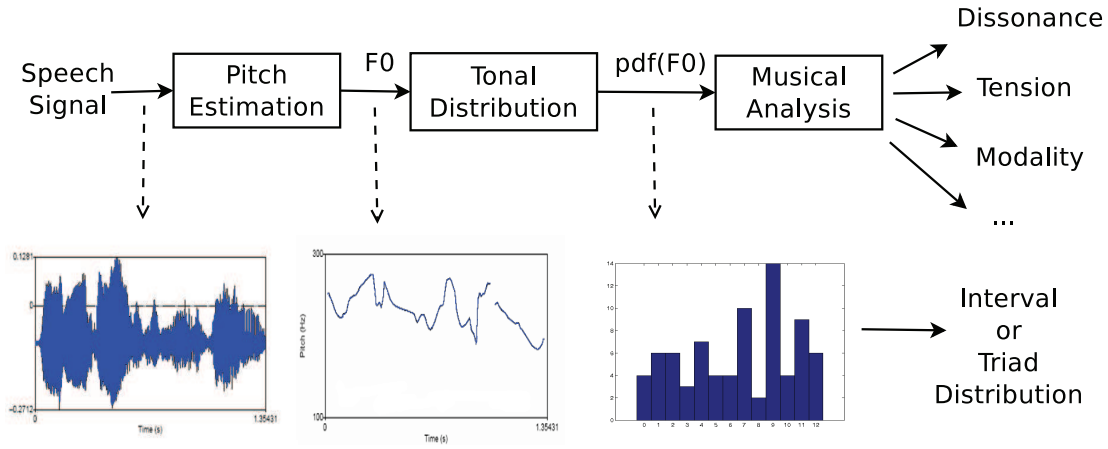


Figure 5.1.: Extraction of tonal distribution features

5.1. Interval features

According to [101], let s be the realization of the random variable pitch, and $p(s)$ its probability density function (PDF)[1]. The distribution $r(s)$ of pitch intervals within an utterance is given by the second-order autocorrelation of the pitch distribution $p(s)$. It can be expressed as:

[1] In our case, $p(s)$ will be approximated with the histogram of all the pitch values.

$$r(s) = \int_{-\infty}^{\infty} p(s+\lambda)\, p(\lambda)\, d\lambda \tag{5.1}$$

The following properties of $r(s)$ hold:

- $r(s)$ is a probability density function, and therefore $\int_{-\infty}^{\infty} r(s)\, ds = 1$.
- If the PDF of the pitch is discrete, that is $p(s) = \sum_i P_i \delta(s - s_i)$, then

$$r(s) = \left(\sum_i P_i^2 \right) \delta(s) + \sum_{i \neq j} P_i P_j \left[\delta(s - s_{ij}) + \delta(s - s_{ji}) \right] \tag{5.2}$$

  with $s_{ij} = s_i - s_j$. The first term is the correlation of each pitch with itself and is not relevant for a harmony study, The remaining mixture terms reflect pitch distances of size $(s_i - s_j)$ or $(s_j - s_i)$.

According to music theory, modifications of pitch frequencies by powers of 2 (octaves) do not affect the perception of pitch intervals (see definition of pitch chroma in Section 3.1.1). We will therefore introduce a circular pitch on the semitone scale:

$$s_o = \mathrm{mod}_L(s) \tag{5.3}$$

That means that the new variable s_o will be in the range $[0, L)$ covering a complete octave. This formula is used to map all octaves into a single one. In the literature, similar concepts like *pitch class profiles* or *chroma vectors* can be found for chord estimation. The difference between these studies and the one performed herein is that they can use directly the short-time spectrum of music signals (harmony in music is vertical) whereas in this case we estimate the pitch, calculate its histogram for the whole sentence and compute its auto-correlation.

The probability density function of s_o can be expressed as

$$p_o(s) = \begin{cases} \sum_{k=-\infty}^{\infty} p(s + kL) & 0 \leq s < L \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

Afterwards, let us compute the autocorrelation of the circular pitch density function $p(s_o)$ like this

$$r_o(s) = \int_0^L p_o(\mathrm{mod}_L(s+\lambda))\, p_o(\lambda)\, d\lambda \tag{5.5}$$

This distribution can be interpreted as the probability density function of the interval distribution within an octave; in other words, $r_o(s)$ is the PDF of the circular pitch distance $I_o = \mathrm{mod}(s_1 - s_2)$. Other interesting properties of $r_o(s)$ are the following:

- $r_o(s) = \sum_{k=-\infty}^{\infty} r(s + kL)$
- $r_o(s) = r_o(L - s)$ This symmetry property explains why two complementary intervals in music with a sum equal to one octave (for example, the perfect 4th and perfect 5th) causes the same consonant or dissonant effect.

In addition to the PDF $r_o(s)$, it is possible to calculate the total intervallic dissonance of the distribution. Let $d(s)$ be a suitable dissonance function for the logarithmic pitch distance s or equivalently the frequency ratio $2^{s/L}$. The more dissonant an interval s_k is, the higher the value of $d(s_k)$ is. Now, let us define the mean dissonance DIS:

$$DIS = \int_0^L d(s)\, r_o(s)\, ds \tag{5.6}$$

This parameter DIS corresponds to the expectation of the random variable $d(I_o)$ where $I_o = \mathrm{mod}_L(s_1 - s_2)$ is the circular pitch distance with PDF $r_o(s)$. In our case, the dissonance function $d(s)$ will be based on frequency ratios[2], corresponding to the geometric mean $\sqrt{N(s)\,D(s)}$ where $N(s)/D(s)$ is the rational approximation of the frequency ratio $2^{2/L}$ with a tolerance value of 0.02 (see Figure 5.2).

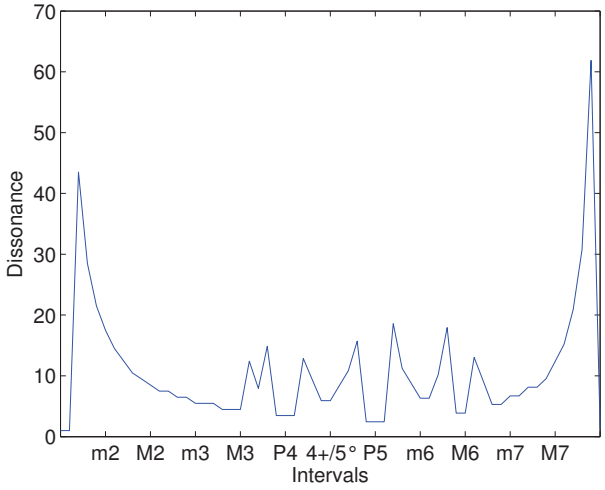


Figure 5.2.: Interval dissonance calculated as the geometric mean

The total number of interval features is 31 from which 30 correspond to the histogram values approximating the distribution $r_o(s)$, the last one being the mean dissonance DIS.

5.2. Autocorrelation triad features

In contrast to intervals, the perception of three-note chords like major, minor, diminished and augmented is not completely understood yet. Nevertheless, the concept of autocorrelation of pitch PDF can easily be extended to triads. The third-order circular autocorrelation of $p_o(s)$ is given by

[2]Each musical interval can be represented by a ratio of two integers N/D. For example a fourth corresponds to 3/4 or a fifth to 2/3.

$$r_o(s_1, s_2) = \int_0^L p_o(\mathrm{mod}_L(s_1 + \lambda))\, p_o(\mathrm{mod}_L(s_2 + \lambda))\, p_o(\lambda)\, d\lambda \tag{5.7}$$

Figure 5.3 shows the triad distribution for a particular speech utterance.

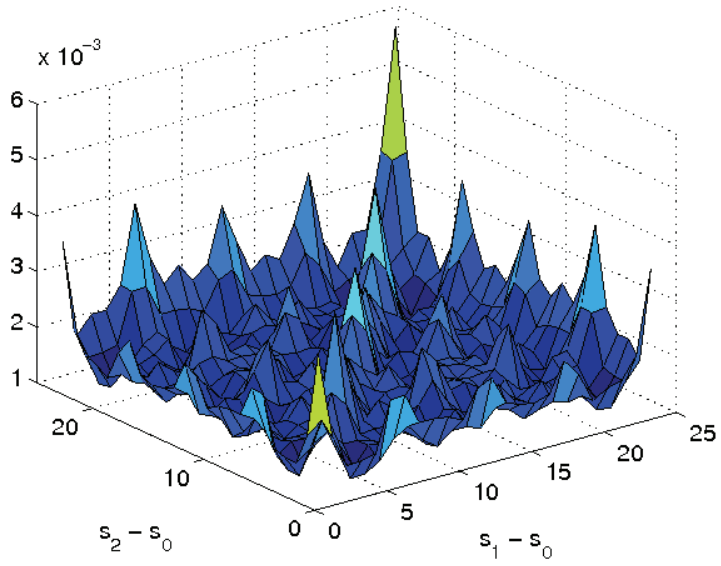


Figure 5.3.: Example of the third-order autocorrelation of the circular pitch

There are three different possibilities to generate features from this plot:

1. Give all the values of the distribution to the classifier directly.
2. Look for maxima in the plot and give their position and intensity values as features.
3. Take specific significant points of the plot. For example, points corresponding to important triads like major, minor, diminished or augmented.

The first option would produce a big amount of features, which would entail an important computational cost in the feature selection step. The second option is also very expensive computationally speaking in order to find the maxima, but it is definitely a good future path. In our case, we have opted for the last option: this solution is not optimal in the sense that we might be discarding important information, but it is simple and fast, and correspond to special musical triads. In our implementation, we generated the points corresponding to the four most famous triads: major, minor, diminished and augmented.

5.3. Gaussian triad features

In the previous section, we have obtained the complete distribution of triads, taking into account all the pitch information. Another possibility would be to consider only

the most important pitch values to calculate the triad distribution. The idea here is to extract the principal musical components or essential harmonic structure, discarding all the small pitch values as noise [13].

We are going to apply the psychophysical model of triad perception explained in Section 3.3 to the analysis of speech intonation. First of all dominant pitches have to be selected. Given a certain interval distribution, we try to fit the minimum number of gaussian distributions that would give the best approximation. Figure 5.4 shows an example of gaussian fitting for the histogram of a certain pitch contour.

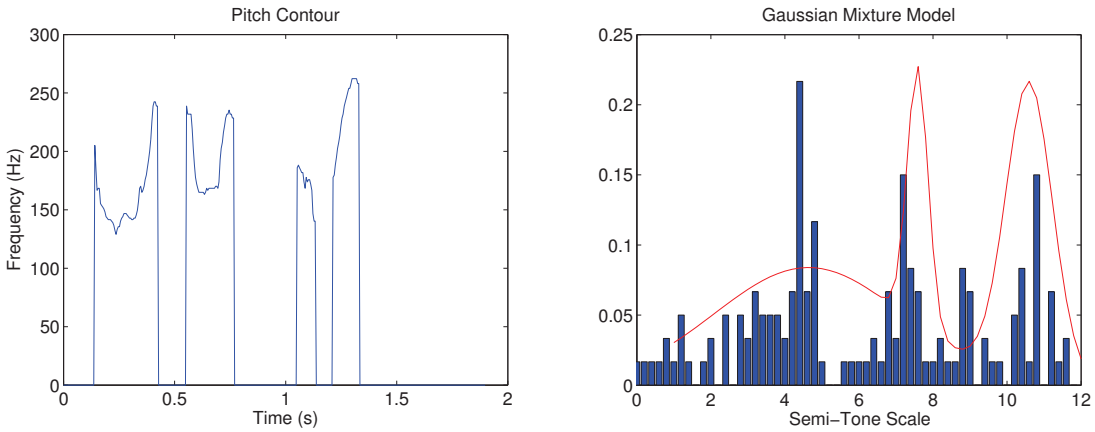


Figure 5.4.: Extraction of dominant pitches for a happy utterance

The problem of finding "clusters" or gaussian mixtures corresponds to an unsupervised learning problem which can be solved with an expectation-maximization EM algorithm and the minimum description length MDL order estimation criterion. The number of clusters can be directly derived from the data with the help of the Akaike information criterion AIC [4]. This criterion measures the fit goodness of a statistical model. It takes into account accuracy and simplicity of the model, having a penalty term for complex models. The AIC can be expressed as

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{5.8}$$

where k is the number of parameters in the statistical model, and L refers to the maximized value of the likelihood function for the estimated model. The preferred model will be the one with minimum AIC value. Once the principal pitch values have been found, musical attributes like dissonance, tension and modality of the chord can be calculated with the formulae presented in Section 3.3.

Additionally, features involving the parameters of the Gaussian fit directly can be generated. Following [24], the gaussian with the highest amplitude a_{ton} will be considered the *tonic mode* (tone at the bottom of the triad) at position k_{ton} and standard deviation σ_{ton}. The minimum and median mixtures have amplitudes a_{min} and a_{med}, means at k_{min} and k_{med} and standard deviations σ_{min} and σ_{med} respectively.

The following features can be defined:

$$\pi = a_{ton} \tag{5.9}$$
$$\pi_{min} = \frac{a_{min}}{a_{ton}} \tag{5.10}$$
$$\pi_{med} = \frac{a_{med}}{a_{ton}} \tag{5.11}$$
$$s_{min} = \frac{\log \sigma_{min}}{\log \sigma_{ton}} \tag{5.12}$$
$$s_{med} = \frac{\log \sigma_{med}}{\log \sigma_{ton}} \tag{5.13}$$
$$FR_1 = \frac{k_{med}}{k_{ton}} \tag{5.14}$$
$$FR_2 = \frac{k_{min}}{k_{med}} \tag{5.15}$$

The values π_{min} and π_{med} account for the strength of the weakest and median tones with respect to the tonic mode. s_{min} and s_{med} measure how concentrated or spread each constituent of the chord is in respect to the tonic mode. Finally, FR_1 and FR_2 account for the intervals in terms of ratios. These features represent the trimodal structure of an F0 signal.

For both interval and triad features, temporal information is lost. The variation of pitch along time is not being taken into account, only the cumulative histograms are considered. In the following section, we will try to create other features based on a perceptual model of intonation [35][56].

5.4. Perceptual model of intonation

According to what was said in Section 2.3 on page 16, pitch modulation serves multiple prosodic functions, making even a neutral sentence rich in terms of harmony. Our hypothesis is that emotions are not uniformly expressed in time; emotional content rather arises at specific points in time with more or less strength. When performing pitch distributional analysis, the whole utterance is examined, even if emotional information might be present only in a small fraction of it.

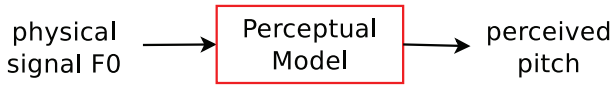


Figure 5.5.: Approximation of pitch perception

In this section, we ignore all the segments with "no-emotional speech" by applying syllabic segmentation and stylization algorithms on the F0 signal. Musical analysis on speech shall give better results if we focus on the *perceived pitch* rather than directly on the F0 contour. In other words, we will simplify the physical signal F0 into a pitch

signal closer to the one that reaches our brain (see Figure 5.5). For this, let us enquire into psychoacoustic properties of pitch perception [35].

5.4.1. Perceptual principles

Pitch Perception is based on a central processor which detects the Greatest Common Divisor between harmonic frequencies (for example, the result for the sequence 300, 450 and 750 Hz would be 150Hz). This processor can work with very short-time signals (30ms) and only with two harmonics; it is furthermore tolerant to big intensity differences between them (up to 25 dB) and operates in the range between 50 and 500Hz.

We may wonder how the physical signal F0 is processed inside our brain. How accurately can we perceive it? Section 3.1 on page 22 already gave some insight on the matter of psychoacoustics. We explained the mechanism of our auditory system as a filter bank logarithmically distributed. Here, we shall expose additional psychoacoustic concepts related to the following four perceptual thresholds:

1. Differential Threshold Pitch or Just Noticeable Difference (JND): smallest perceivable difference in frequency between two successive tones. Quite accurate, this threshold increases for very small and very big frequencies as well as with smaller durations below 30ms. In the range of F0, the JND threshold is around 3 Hz.

2. Differential Threshold Pitch Distance: conditions in which a listener is able to judge one interval bigger or smaller compared to another one. In this case, the threshold is around 1.5 or 2 semitones (ST). Roughly speaking, only differences of more than 3 ST can play a part in communicative situations. For instance, if we have a rise between 125Hz and 150Hz (3.16 ST), then another rise starting at 100Hz has to reach at least 143Hz (6.19 ST) in order to be perceived as bigger.

3. Glissando Threshold or Absolute Threshold of Pitch Change: A change in pitch requires a minimal amount of frequency change, otherwise it is perceived as a static pitch. This threshold g_{th} measures how much F0 should change (given a certain interval of time) in order to evoke a sensation of pitch change; it can be expressed with the following formula:

$$g_{th} = 0.16/T^2 \ \left[\mathrm{ST/s^2}\right] \tag{5.16}$$

   where T is the duration of the pitch change in seconds.

4. Differential Glissando Threshold or differential threshold of Pitch Change: In the case of two successive glides, there is a minimum difference between slopes in order not to be perceived as a single glide. This threshold dg_{th} is proportional to the slopes a_1 and a_2 and depends on the duration of glides. Depending on contextual conditions, it lies on a range between 1 and 30 ST/s. In our implementation, we will take:

$$dg_{th} = a_2 - a_1 = 20 \ [\mathrm{ST/s}]$$

This is the biggest perceptual threshold, which means that the differential sensitivity to rate of frequency change is rather low. In other words, we tend to aggregate multiple slopes in a single one if the difference in slopes is not big enough.

5.4.2. Pitch perception in real speech

Pitch variations in real speech come along with formant and amplitude variations, which makes pitch perception even harder. The same F0 movement can indeed be perceived as a pitch glide or two level tones depending on the segmental context. Pitch variations are generally better perceived when followed by a pause. The final parts of the segments tend to be perceived better: this is the so-called perceptual "2/3 rule", which takes as representative tone of the segment the pitch level located at two thirds of the segment.

In the case of vowels with interleaved consonants $V_1CV_2CV_3$, the differential glissando threshold can be strongly damaged. Even more critically, a drop in amplitude between 10 and 20 dB (often found in transitions from vowels to consonants) can obscure changes in F0 of up to half an octave. Therefore, perceptual experiments with entire speech utterances have still to be held.

5.5. Syllabic segmentation algorithm

Melody in speech tends to be perceived as discrete tones associated with syllabic events. Therefore, it would be interesting to have a syllabic segmentation of our speech signals based on their acoustical properties. This could be useful for stylization algorithms, and help to give a better approximation of the perceived pitch.

The syllabic segmentation algorithm is based on [24] and uses both spectral and energy information. First, acoustic segments are identified using the blind segmentation algorithm, identifying the boundaries in the spectral image of the signal. Afterwards, the acoustic segmentation is refined by localizing the syllabic nuclei boundaries, corresponding to energy peaks.

5.5.1. Spectral segmentation

The objective of this algorithm is to segment the speech signal according to its spectral variations. An overview of the successive steps is given in Algorithm 5.1.

We call an *acoustic segment* (simple or complex) a region of stationary spectral content. In order to detect those segments, we will tract the temporal spectral variability by building an Inter-Frame Variability Function IFV, defined as:

$$IFV_d(k) = \sum_{r=k-1}^{k} \sum_{s=k+1}^{k+2} d\left(X(r, m), X(s, m)\right) \tag{5.17}$$

where $X(k, m)$ denotes the squared magnitude of the short-time Fourier transform, m is the frame index and k the frequency index. The highest variability in terms of spectral content (maxima in the IFV function) will correspond to segmental transitions.

Algorithm 5.1 Spectral segmentation algorithm

1. Short-time Fourier transform
2. For each frequency band, apply Canny's algorithm in order to find local vertical edges in the spectrum.
3. Compute the IFV function by averaging over local vertical edges
4. Find peaks in the IFV function

The idea is to detect local vertical edges in the spectrogram of the speech signal. For this, edge identification for image processing is used by applying Canny's algorithm, described in 5.2. More details can be found in [24].

The analysis is held independently for each frequency band and then all the votes across frequencies are added. Sharp transitions will differ in more frequency bands, which will attribute more weight to the IFV function. An example of IFV function and speech segmentation is presented in Figure 5.6.

Algorithm 5.2 Canny's algorithm for edge detection

1. Gradient estimation (convolution with two operators: Gaussian and first derivative of a Gaussian)
2. Non-maximum suppression (linear interpolation around each pixel to determine maxima for some direction in the image)
3. Hysteresis thresholding: local resistance to noise in regions of small spectral change

5.5.2. Identification of syllable nuclei

It is well known that vowels are strongly periodic signals associated to high energy in the first formants. Syllable nuclei will hence correspond to moments of high energy values in the low band of the spectrum. The algorithm for syllable nuclei localization is very simple and is described in Algorithm 5.3.

5.6. Stylization algorithm

A stylization algorithm consists in modifying the measured F0 contour of an utterance into a simplified but functionally equivalent form, preserving all the melodic information useful for communication [54][16]. Not only does it reduce the data by filtering irrelevant F0 events, but it also allows a better analysis of macroscopic events. It basically attempts to simulate tonal perception (what is actually heard).

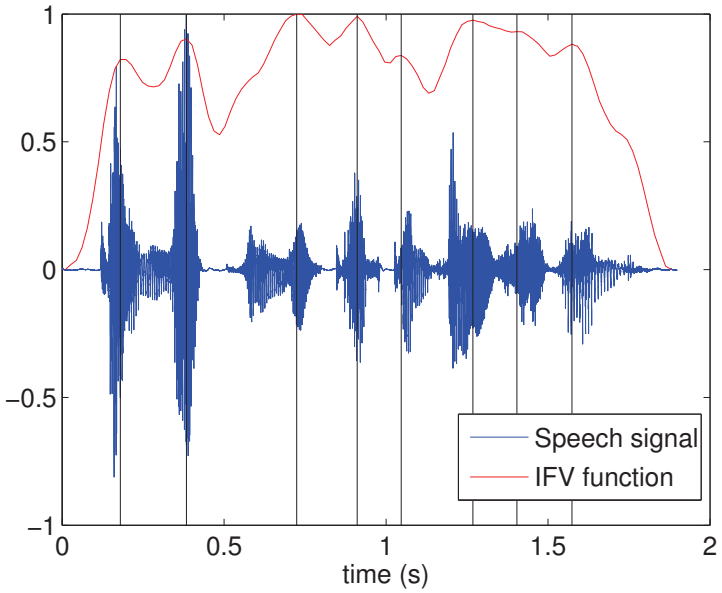


Figure 5.6.: Segmented speech signal using spectral information

Algorithm 5.3 Syllable nuclei localization algorithm

1. Band-pass filter the speech signal (Butterworth)
2. Compute the short-term energy
3. Find peaks p_k and valleys v_k in the energy signal
4. Find the largest normalized jump and compare it to a threshold: discard small peaks and keep the others as the location of syllable nuclei

The implemented algorithm based on [16] will take into account two of the previously seen perceptual thresholds, namely the Glissando Threshold and Differential Glissando Threshold (see Section 5.4.1). Syllabic segmentation is performed in order to reflect that we tend to perceive short *discrete tonal events* aligned with the syllables rather than a continuous pitch contour. Finally, short-term integration simulates the fact that our auditory system is unable to follow rapid changes in F0. Final parts of the tone have furthermore larger weight on pitch judgment than initial ones. In order to simulate these observations, a Weighted Time Average WTA model has been proposed in [16], given by

$$p\left(t\right) = \frac{\int_0^t e^{-\alpha(t-\tau)} f\left(\tau\right) d\tau}{\int_0^t e^{-\alpha(t-\tau)} d\tau}$$

where α accounts for weighting of the past.

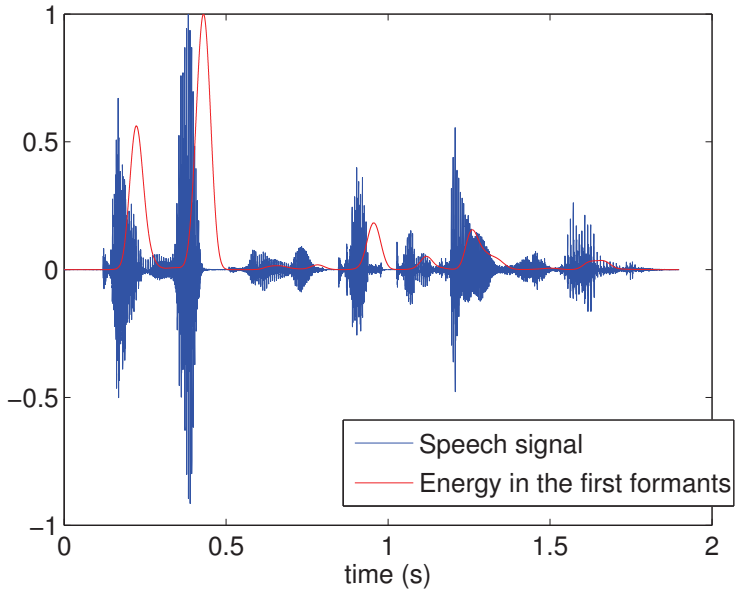


Figure 5.7.: Syllable nuclei corresponding to low-band energy peaks

Figure 5.8 shows an example of pitch stylization. Stylization 1 refers to the implementation explained in Algorithm 5.4. Stylization 2 is another implementation based on [24] that we will not explain.

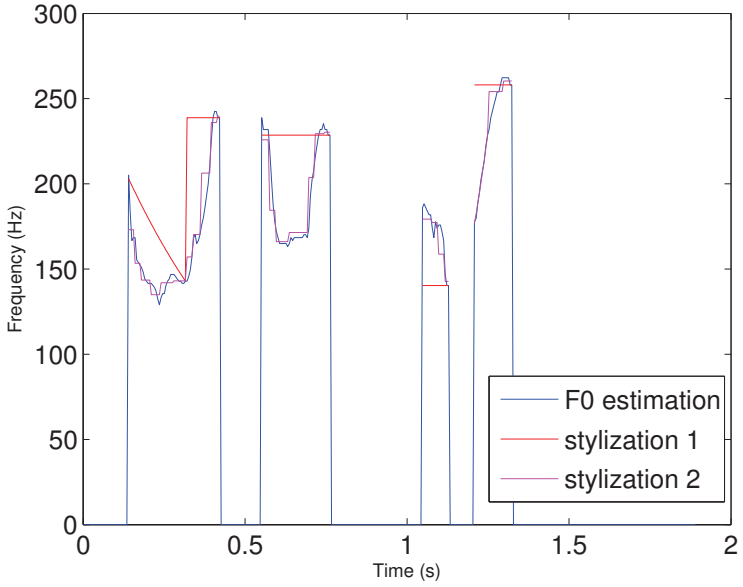


Figure 5.8.: Example of pitch stylization

Algorithm 5.4 Stylization algorithm

1. Speech segmentation (separation in voiced and unvoiced parts)
2. Short-term perceptual integration (applying a Weighted Time Average WTA model to smooth the pitch contour)
3. Syllabic contour segmentation: one or more tonal segments per syllable
   - 1st step (recursive): location of turning points in the contour to break it down into distinct tonal segments
   - 2nd step: decision as to which candidate segments should be aggregated
4. Actual stylization: Pitch Targets are obtained as the pitch at the start and end of the tonal segment. For static tonal segments, the pitch corresponds to the final pitch.

5.7. Features for music emotion recognition

In order to derive new musical features for speech, it seems sensible to look at the features used for emotion recognition in music signals [42][49][103][60][45][95]. The most common features in this field have been listed in Table 5.1. Dynamics, global statistics concerning pitch, as well as MFCCs (defined in Section 4.4 on page 39) have already been implemented in the basic set of features. Features concerning harmony and register have indirectly been approached in the previous sections (autocorrelation triad features are analogous to the idea of chronogram-based features).

Type	Features
Dynamics	Energy
Pitch	Global statistics (mean, range, standard deviation...)
Timbre	MFCCs, spectral shape, spectral contrast
Loudness	intensity, perceptual approaches
Harmony	Roughness, harmonic change, key clarity, majorness
Register	Chromagram, chroma centroid and deviation
Rhythm	Rhythm strength, regularity, tempo
Articulation	Event density, attack slope, attack time

Table 5.1.: Most common features for emotion recognition in music

In the following, we are going to describe three different sets of features concerning loudness, timbre and rhythm [51][50].

5.7.1. Intensity

Intensity features per frame are calculated on the absolute value of the FFT spectrum. They refer to the spectral sum of the signal and spectral distribution in each sub-band, given by

$$I(k) = \sum_{n=0}^{N/2} |\mathrm{FFT}_k(n)| \tag{5.18}$$

$$D_i(k) = \frac{1}{I(k)} \sum_{n=L_i}^{H_i} |\mathrm{FFT}_k(n)| \tag{5.19}$$

where k refers to the frame, $I(k)$ is the intensity of the k-th frame and $D_i(k)$ is the intensity ratio of the i-th subband. L_i and H_i are respectively the lower and upper bounds of the i-th subband.

5.7.2. Timbre

MFCC has proved to be very successful for general speech and music processing, and is therefore often used to describe timbre features. However, it averages the spectral distribution in each subband, and hence loses the relative spectral information. For this reason, new spectral contrast features should be implemented. The computation is done for each subband of the spectrum. Given the FFT vector of the k-th subband $\{x_{k1} \dots x_{kN}\}$, we will sort it in descending order. The resulting vector $\left\{x'_{k1} \dots x'_{kN}\right\}$ will be used to estimate the strength of the spectral peaks and spectral valleys with the following formulae:

$$Peak(k) = \log\left\{\frac{1}{\alpha N} \sum_{i=1}^{\alpha N} x'_{ki}\right\} \tag{5.20}$$

$$Valley(k) = \log\left\{\frac{1}{\alpha N} \sum_{i=1}^{\alpha N} x'_{k(N-i+1)}\right\} \tag{5.21}$$

where α is used as a small neighborhood factor and set to 0.2. The spectral peaks and spectral valleys are estimated using the mean of the largest and lowest values in the spectrum, respectively, instead of the exact maximum and minimum values. The corresponding spectral contrast is given by

$$SC_k = Peak_k - Valley_k$$

In our case, we will directly give the values $Peak_k$ and $Valley_k$ to the classifier.

5.7.3. Rhythm

Three different aspects of rhythm are closely related to the emotional response to music: rhythm strength, rhythm regularity and tempo [50]. The procedure to calculate these rhythm features is the following:

1. For each frame, compute the FFT.
2. Divide each FFT into seven octave-based subbands.
3. Get the amplitude envelope of each subband by convoluting with a half-Hanning (raised cosine) window
$$A_i^{'}(n) = A_i(n) \otimes h_w(n) \tag{5.22}$$
where $A_i(n)$ is the amplitude of the ith subband, $A_i^{'}(n)$ is the corresponding amplitude envelope and $h_w(n)$ is the half-Hanning window, defined as
$$h_w(n) = 0.5 + 0.5\cos\left(2\pi \cdot \frac{n}{(2L-1)}\right) \quad n \in [0 \ldots (L-1)] \tag{5.23}$$
where L is the length of the window, empirically set to 12.
The half-Hanning window is generally used for envelope extraction while keeping sharp attacks in the amplitude curve.

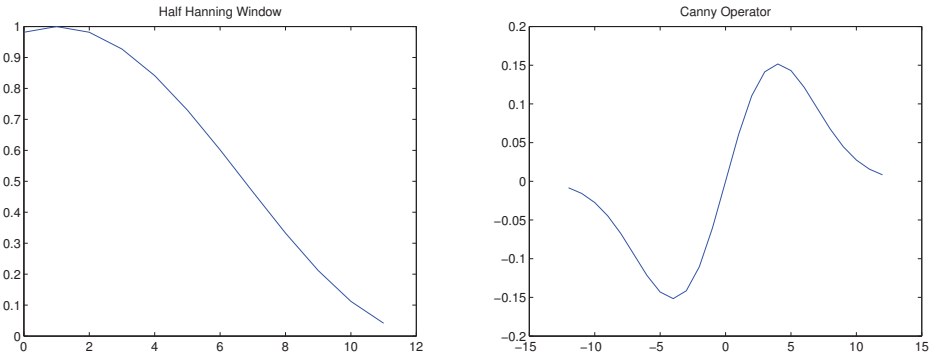


Figure 5.9.: Operators used for rhythm extraction

4. Apply the Canny operator to detect onset sequences, i.e., how the amplitude envelope varies.
$$O_i(n) = A_i^{'}(n) \otimes C(n) \tag{5.24}$$
where $O_i(n)$ is the onset sequence of the i-th subband and $C(n)$ is the Canny operator with a Gaussian kernel, defined as
$$C(n) = \frac{n}{\sigma^2} e^{-\frac{n^2}{2\sigma^2}} \quad n \in [-L_c \ldots L_c] \tag{5.25}$$
The parameter σ is used to control the operator's shape, and L_c refers to its length. The Canny operator is usually used for edge detection in image processing.

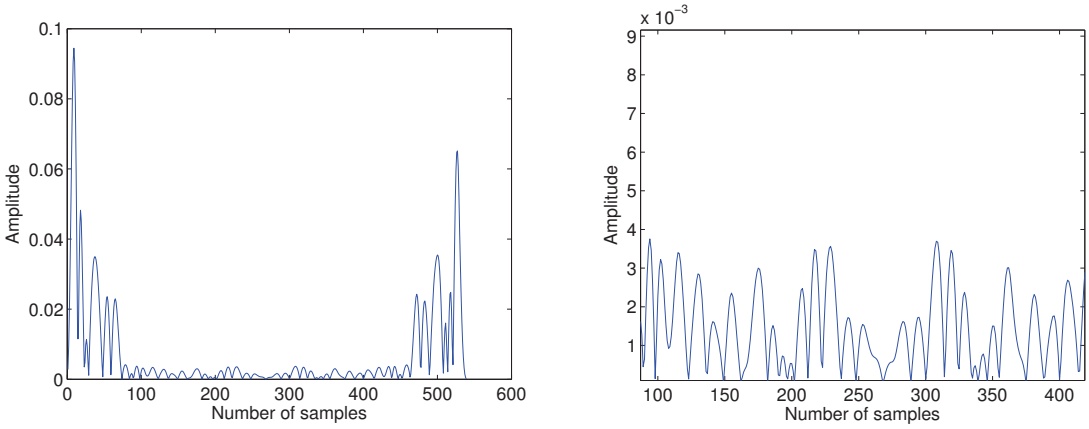
5. Sum up the onset curve of each subband. The obtained curve, called Onset sequence, is used to represent the rhythm information of the signal (see examples in Figure 5.10).

6. Extract the rhythm features:

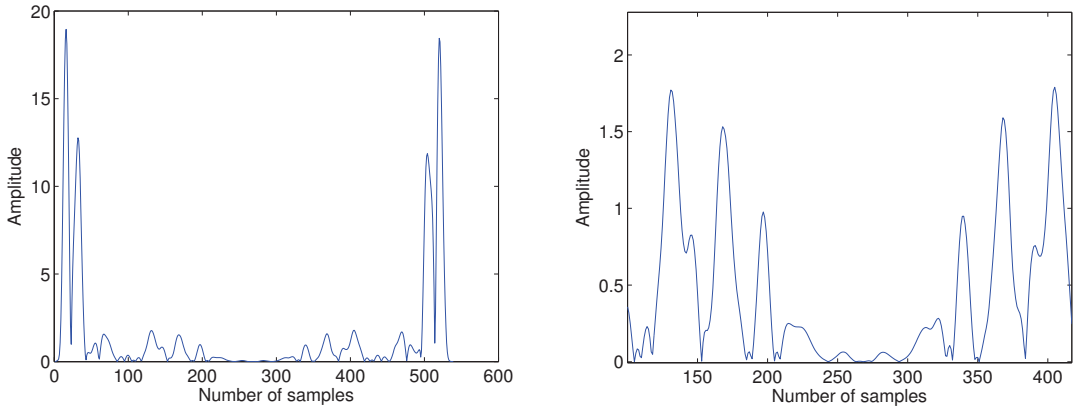
   Rhythm strength Average onset strength in the onset curve; in other words, average value of the peaks of the onset curve. The stronger the rhythm is, the higher the value.

   Rhythm regularity Average strength of the local peaks in the autocorrelation of the onset curve. The higher the value is, the more regular the rhythm.

   Rhythm speed Ratio of number of peaks in the onset curve and the corresponding time duration. The more peaks we have for a given duration, the faster the rhythm is.

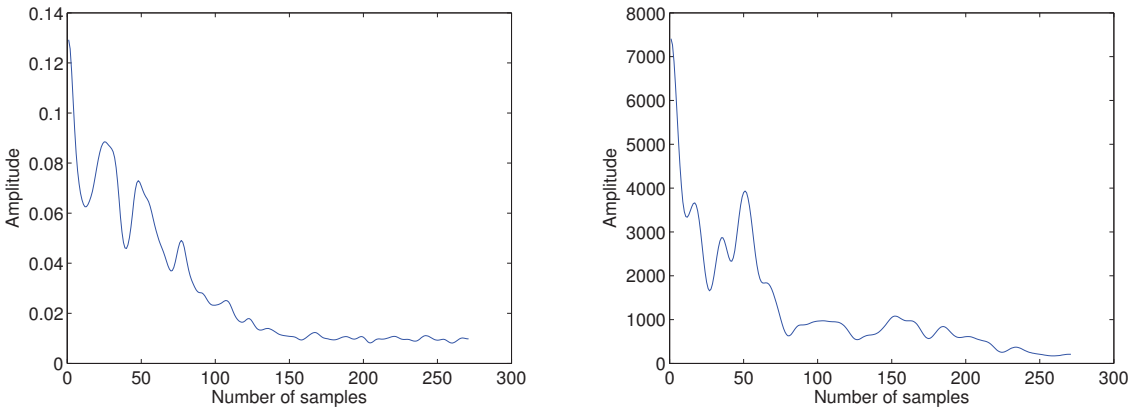


(a) Onset sequence 1



(b) Onset sequence 2

Figure 5.10.: Examples of onset sequences



(a) Autocorrelation 1

(b) Autocorrelation 2

Figure 5.11.: Examples of autocorrelation for their corresponding onset sequences

5.8. Conclusion

In this section, we have presented three families of musical features. The first one is based directly on music theory (interval and triad features). The second one, based on linguistics, applies an stylization algorithm to the F0 signal before computing the features of the first family. Finally, the third category of features have been borrowed from the field of emotion recognition from music signals, and concern three important characteristics of music: loudness, timbre and rhythm. In the next section, we will investigate how much these musical features can improve the recognition rate.

6. Simulations and Results

6.1. Emotional speech database

Finding a suitable emotional speech database is a very complicated task. Very few and sparse public databases are actually available. Our experiments have been held on acted emotions (rather than natural or induced emotional speech). We have used two different emotional speech databases: the TUB (Technische Universität Berlin) emotional speech database and the SES (Spanish Emotional Speech) database. Their characteristics have been summarized in Table 6.1.

Since all the utterances in the SES database correspond to a single speaker, the majority of our experiments have been held on the TUB database for a speaker-independent classification.

Description of the Berlin Emotional Database

The TUB database consists of seven different emotions, but we have decided to remove the "disgusted" emotion in order to be able to apply hierarchic classification, based on the emotional model of Figure 1.3 on page 9. A total of 10 German sentences of emotionally neutral content have been acted by 10 professional actors, 5 of them being female. The database is recorded in 16 bit, 16 kHz under studio noise conditions. Throughout perception tests performed by 20 evaluation listeners, 488 audio files have been proven to be particularly good samples: they have been correctly recognized by 80% of the listeners, and rated as natural by more than 50%. The distribution of the small TUB database has been plotted in Figure 6.10.

	Language	Speakers	Utterances	Sentences	Emotions	Emotion Labels
TUB	German	10	814	10	7	neutral, happy, sad, bored, angry, afraid, disgusted
SES	Spanish	1	165	15	5	neutral, happy, sad, angry, surprised

Table 6.1.: Description of the available databases

big TUB	Neutral	Happy	Sad	Angry	Afraid	Bored	Total
# Samples	105	113	122	138	123	113	708
Percent (%)	14.8	16.0	17.2	19.5	17.4	16.0	100

small TUB	Neutral	Happy	Sad	Angry	Afraid	Bored	Total
# Samples	79	71	62	127	69	81	488
Percent (%)	16.2	14.6	12.7	26.0	14.1	16.7	100

Table 6.2.: Number of samples for each emotion in the TUB database

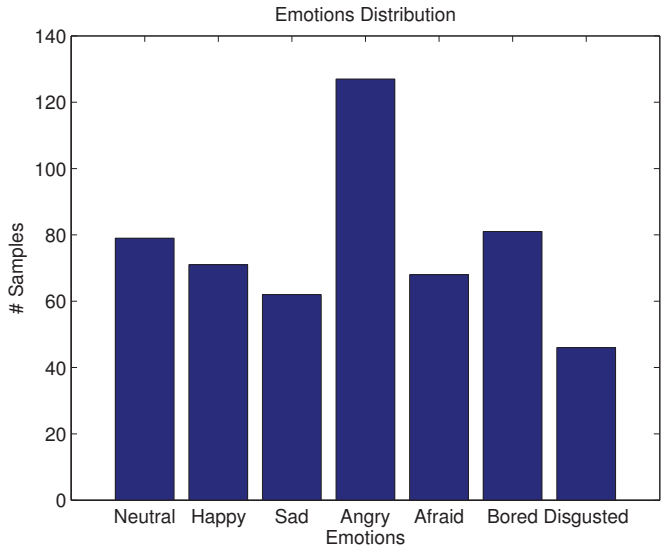


Figure 6.1.: Distribution of the small TUB database (488 files without disgusted)

6.2. Feature sets

In our experiments, we always talk about *basic* and *full set*. Basic set refers to the traditional features described in Section 4.4 except the formant and voice quality sets. Full set corresponds to the basic set combined with the musical features. Tables 6.8 and 6.4 show the nature of the features in the basic and musical sets in our implementation. In our simulations, we have decided not to include the timbre features. Although they improve the results a little bit, it does not compensate the additional computational cost.

duration	MFCCs	ZCR	harmony	energy	pitch	TOTAL
16	91	13	3	58	33	214

Table 6.3.: Basic set of features

interval	auto-correlation triad	gaussian triad	intensity	rhythm	TOTAL
31	4	10	63	15	123

Table 6.4.: Musical set of features

Table 6.5 exposes the family of features used in the old basic set. Finally, the old musical set refers to the interval and autocorrelation triad features.

duration	ZCR	harmony	energy	pitch	formant	voice quality	TOTAL
16	17	3	81	49	35	99	300

Table 6.5.: Old basic set of features

The new architecture separating local and global features not only allows us to select individual sets of features, but also gives faster results, reaching a speed-up above 50%. The slowest functions in the feature extraction step are the RAPT pitch estimation algorithm and VAD (voice activity detection) algorithm.

Figure 6.2 visualizes our data points for four emotions (“happy”, “sad”, “bored” and “angry”) in a simplified two-dimensional space. This space is built by taking the first two components of a Principal Component Analysis (PCA). These directions correspond to the ones of maximal variance, which allows for a good visualization.

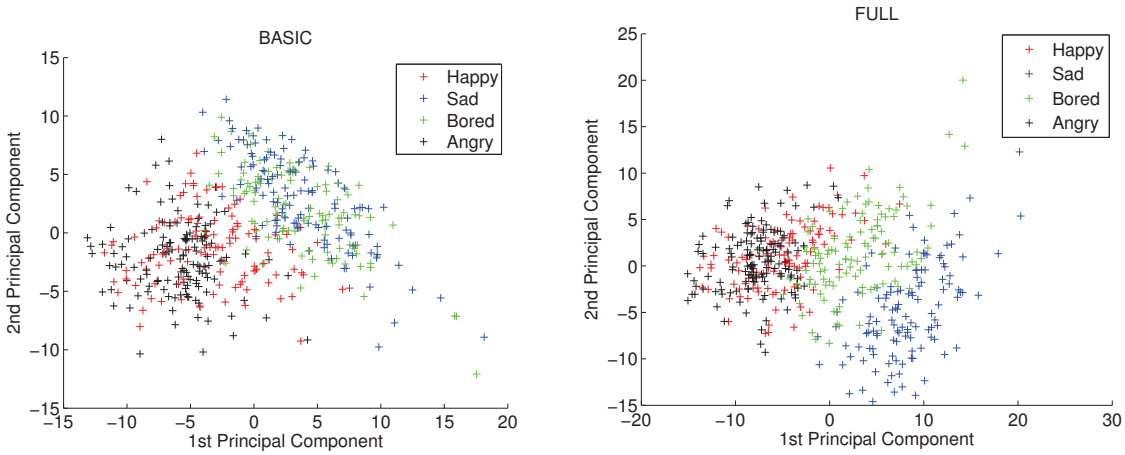


Figure 6.2.: Data visualization, two principal components for the basic and full set of features

In the left side, we can observe that the emotions “sad” and “bored”, as well as “angry” and “happy”, overlap. These two problems might be difficult to solve. In the right side, the addition of musical features has simplified the problem of “bored” versus “sad” (less

overlapping). However, the clusters for "happy" and "angry" are still too close to each other. "Angry" versus "happy" has indeed proved to be the most difficult problem in emotion recognition from speech signal [84].

As a last remark, we would say that they might be other directions (different from the PCA components) in which "happy" and "angry" are better separated. These plots illustrate the improvement in the "sad" versus "bored" classification problem.

6.3. Configuration 9-1 versus 8-1-1

In our implementation, we can follow two different evaluation strategies, which have been illustrated in Figure 6.3. In the first 9-1 configuration, the SFFS feature selection algorithm uses information from all the speakers. A 9-1 loop (leaving one speaker out) is performed inside this algorithm. The training in this case is performed only once (the selected features are the best ones for all the speakers). In the case of 8-1-1 evaluation, training is performed for each speaker independently. We will select a new set of features for each one: now SFFS uses information from only 9 speakers and it will be called 10 times.

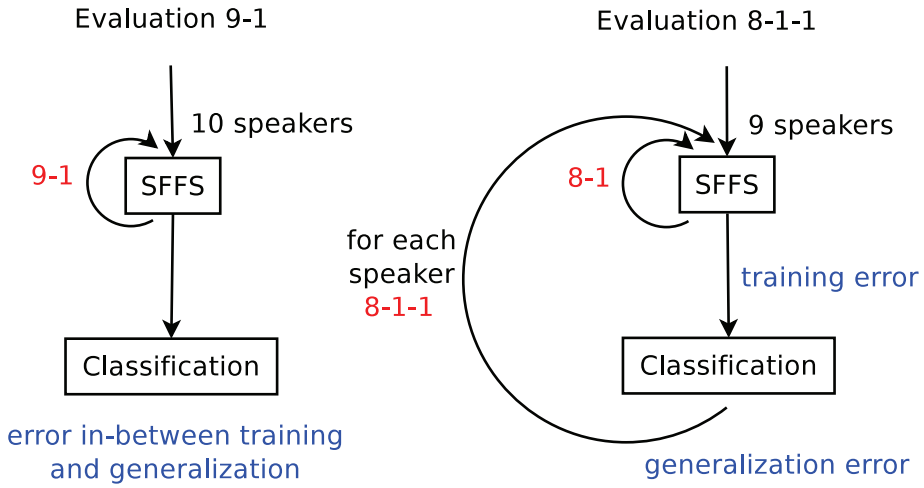


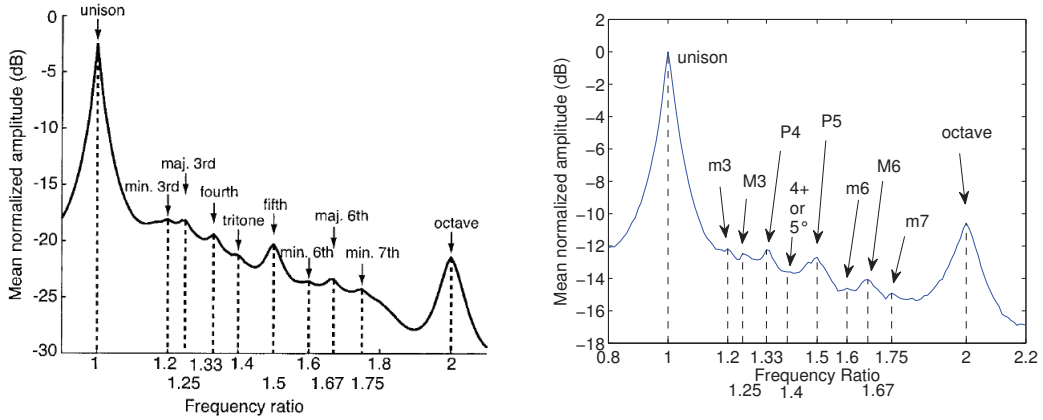
Figure 6.3.: Different strategies for evaluation

8-1-1 evaluation should always be used. Indeed, the 9-1 configuration is not completely correct since information of the speaker to evaluate is used in the training phase; nevertheless, 9-1 evaluation is roughly 10 times faster than the 8-1-1 configuration and allows us to compare different feature sets or classification methods.

6.4. Validation of musical universals

In this experiment, we have reproduced the results of [87], described in Section 2.6. The average FFT is computed for several 100ms blocks, and then normalized both in

amplitude and frequency, against the fundamental frequency F0. The results are plotted in Figure 6.5. The original experiment is represented on the left side; our results for the TUB database can be seen on the right. The fact that both curves present the same peaks confirm that the statistical structure of speech is indeed closely related to musical universals.



(a) Results for the Multi-language Telephone Speech Corpus

(b) Results for the TUB database

Figure 6.4.: Average normalized spectrum of 100ms speech blocks

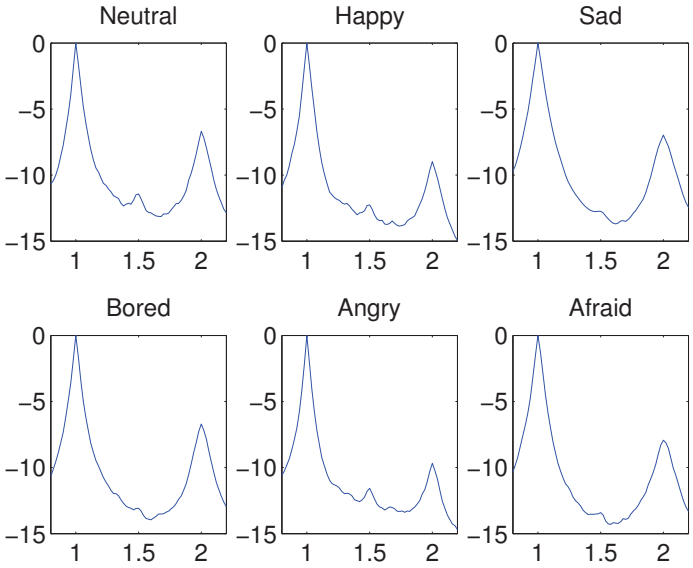


Figure 6.5.: Average normalized spectrum per emotional class

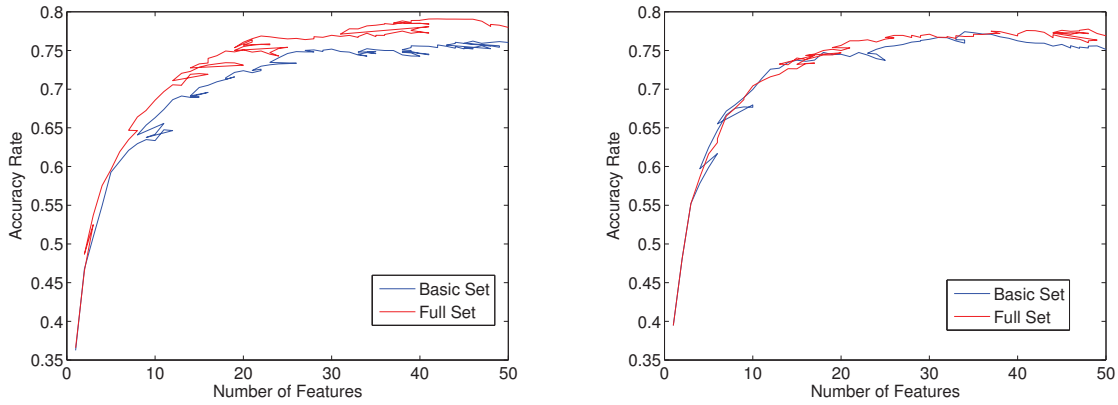
Additionally, the same experiment has been plotted for each emotional class separately in Figure 6.5. In these subplots, peaks at particular frequency ratios are generally less pronounced than in the global plot. The number of samples for each emotion is not sufficient to draw final conclusions (see Table 6.6). However, there is a certain tendency for the low-activation emotions ("sad", "bored" and "afraid") to have a flater curve between the unison and the octave.

Neutral	Happy	Sad	Bored	Angry	Afraid
334	413	442	690	1056	638

Table 6.6.: Number of 100 ms averaged speech blocks for each emotion

6.5. Comparison between old and new features

In this section, we will compare the relative performance of the old and new features. In both cases, 9-1 and 8-1-1 evaluations are performed. The SFFS feature extraction algorithm will select the 50 or 100 best features. In the following, we will see several curves "going backwards": this phenomenon is due to the iterative nature of the SFFS algorithm. The curves are indeed parametrized by the number of iterations.



(a) New implementation

(b) Old implementation

Figure 6.6.: 9-1 evaluation for the plain Bayes classifier

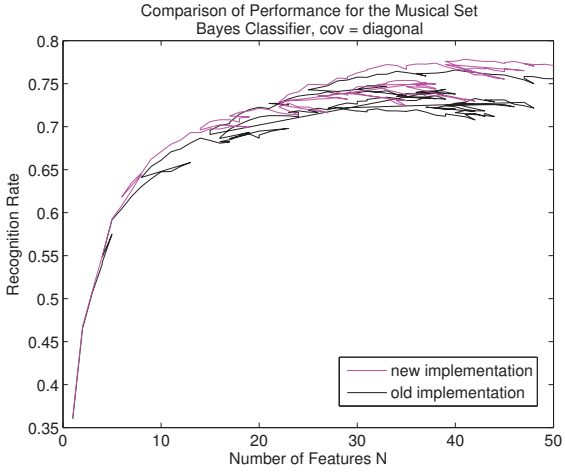


Figure 6.7.: Comparison of the old and new musical features (only interval and triad features) for the plain Bayes classifier

In both types of evaluation, it can be observed that the old basic set of features is slightly better than the new one. Even so, we decided to keep our new basic set because of the following reasons:

- Faster feature generation
- Modularity concerning the different subsets (for example, it is possible to remove the energy family)
- More robust features: the old features have more lacking information (not-a-number positions)

On the other hand, the new musical set gives better results than the old one.

6.6. Comparison of the basic and full sets of features

In Figures 6.8 and 6.9, it can be seen that musical features improve our performance in approximately 2%. Tables 6.8 and 6.9 shows the confusion matrix for a 9-1 evaluation. The accuracy rate for “sad”, “bored”, “angry” and “afraid” have improved, whereas “neutral” and “happy” are worstly recognized.

	Training	Generalization
Basic set	75.94	64.04
Full set	78.09	65.72

Table 6.7.: Accuracy rate for the plain Bayes classifier, new impl., 8-1-1 evaluation

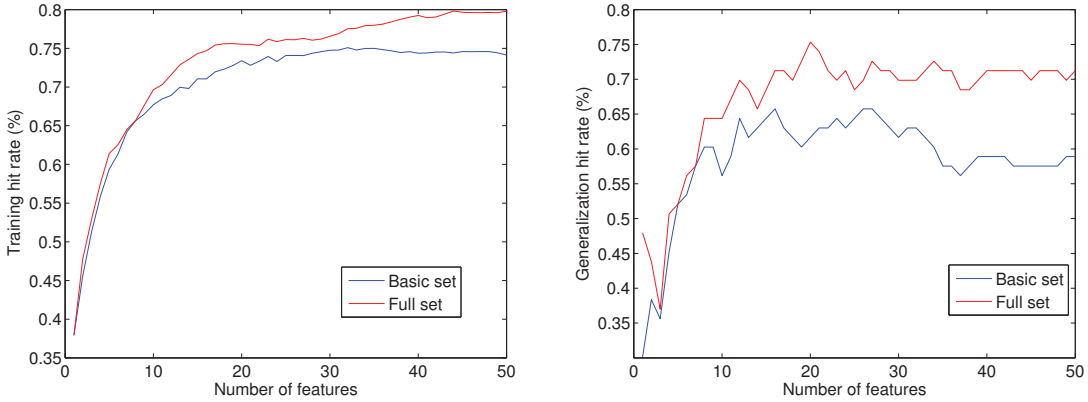


Figure 6.8.: Evaluation with plain Bayes classifier for the first speaker

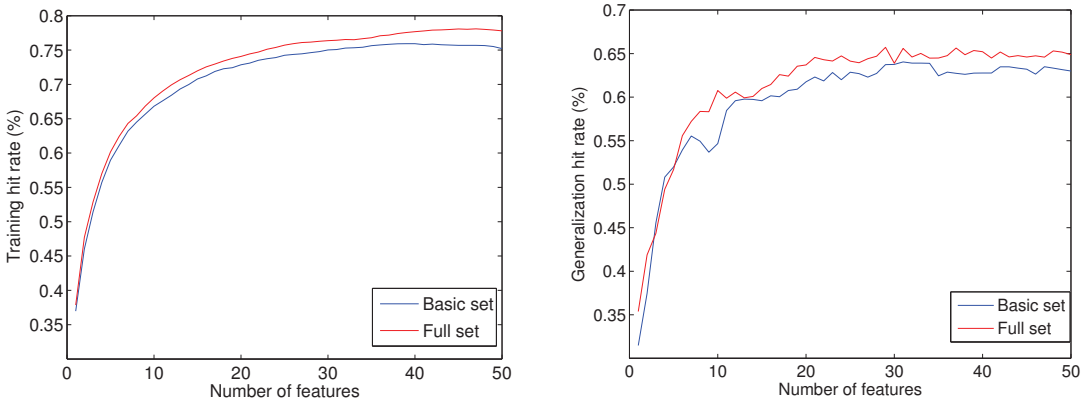


Figure 6.9.: 8-1-1 evaluation for the plain Bayes classifier

	Neutral	Happy	Sad	Bored	Angry	Afraid
Neutral	75.96	1.92	3.85	11.54	1.92	4.81
Happy	3.57	59.82	0.89	2.68	24.11	8.93
Sad	8.26	0	88.43	1.65	0	1.65
Bored	13.39	1.79	3.57	78.57	1.79	0.89
Angry	0	10.95	0	0	84.67	4.38
Afraid	2.46	11.48	2.46	1.64	3.28	78.69

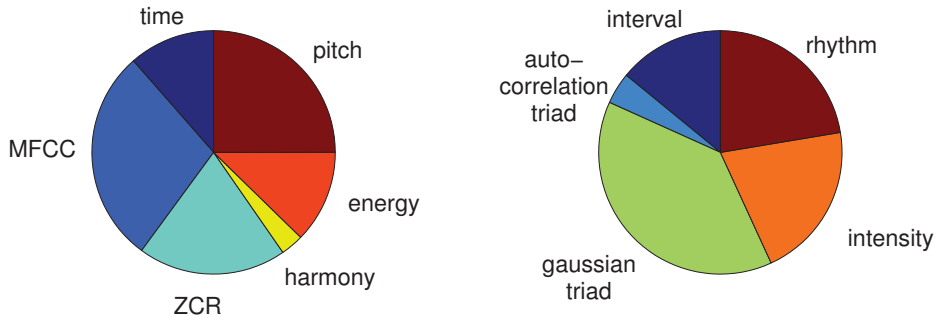
Table 6.9.: Full set of features

	Neutral	Happy	Sad	Bored	Angry	Afraid
Neutral	81.73	1.92	2.88	6.73	0	6.73
Happy	6.25	65.18	0.89	0.89	20.54	6.25
Sad	8.26	0	83.47	6.61	0	1.65
Bored	16.96	1.79	5.36	69.64	1.79	4.46
Angry	0	15.33	0	0	82.48	2.19
Afraid	4.10	10.66	2.46	2.46	5.74	74.59

Table 6.8.: Basic set of features

6.6.1. Analysis of musical features

Figure 6.10 illustrates the relative proportion of features selected by the SFFS algorithm. Around 40% of the selected features where musical ones. Concerning the basic set of features, MFCCs, pitch and ZCR families are often chosen; for the musical features, the gaussian triad are the preferred ones, followed by rhythm and intensity.



(a) Basic Features

(b) Musical features

Figure 6.10.: Distribution of the feature types selected by the SFFS for all speakers

Finally, Figure 6.11 shows the performance for individual musical sets of features. The observations are concordant to the pie graphics of Figure 6.10. The best features are the tonal distribution ones (interval and triad features), followed by rhythm and intensity ones. Features based on the stylization algorithm give the smallest improvement; finally, musical features alone (without the basic set) give accuracy rates around 70%.

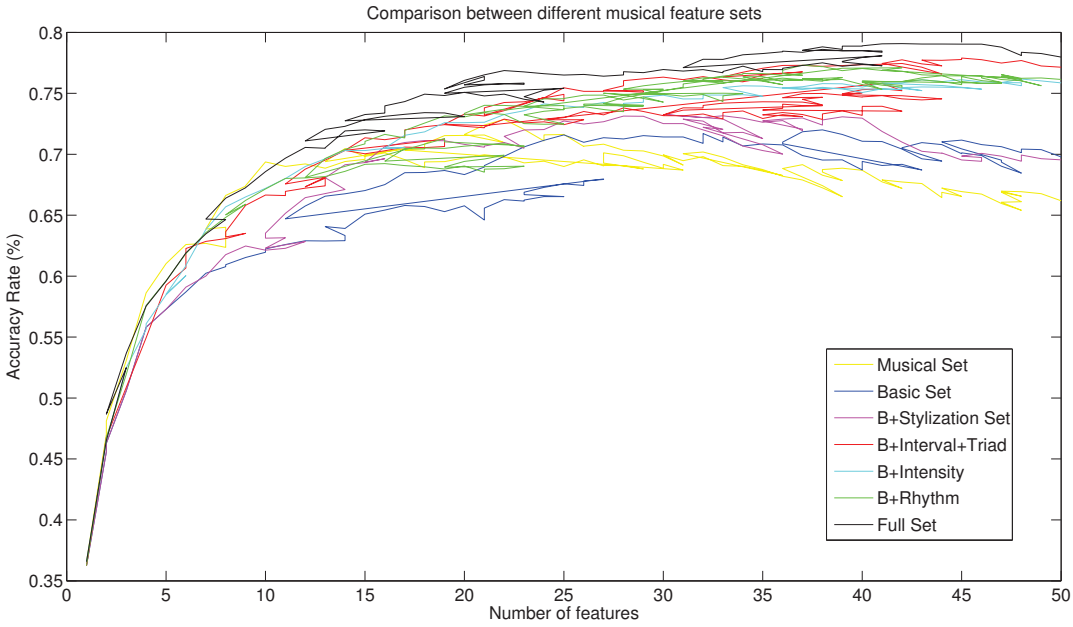


Figure 6.11.: Comparison between different musical feature sets

6.6.2. Comparison between plain and hierarchical Bayes classifier

Table 6.10 shows that a hierarchical classification gives better results than a plain Bayes classifier (around 8.5% better). This makes sense, since we are choosing different sets of features for each binary classification.

	plain Bayes	hierarchical Bayes
Basic	64.04	72.39
Full	65.72	74.16

Table 6.10.: Generalization hit rate for a 8-1-1 evaluation

6.6.3. Happy versus angry

In this experiment, we runned the “happy” versus “angry” classification problem for each speaker. Figure 6.9 shows that musical features improve the accuracy rate in the training phase. In the generalization or evaluation phase, musical features does not seem to help so much. Both cuves for the basic and full set oscillate roughly at same levels (around 15% worse than the training hit rate).

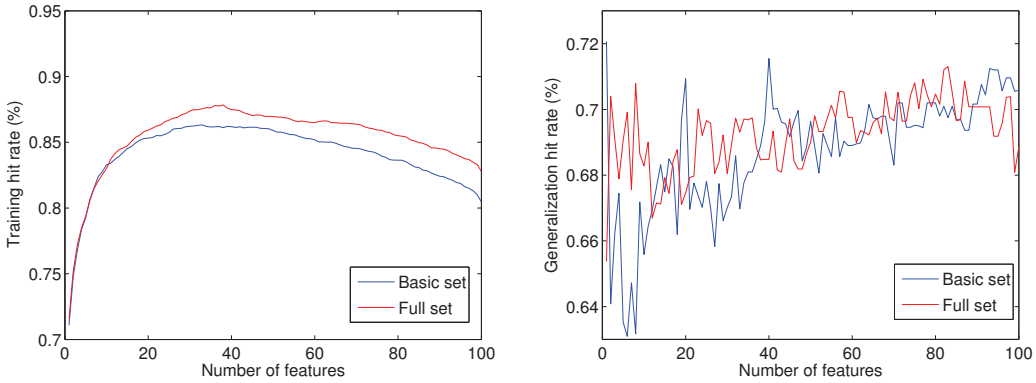


Figure 6.12.: 8-1-1 evaluation for the binary Bayes classifier for angry Vs happy

6.7. Simulations with the SES database

Figure 6.13 shows the results for the Spanish Emotional Speech database. Since this database has utterances for a single speaker, the classification problem is much easier. If we remove the utterances that are very similar (same emotion, same sentence), we get a total of 60 utterances. "Musical45" designates only the interval and triad features, whereas "musicalAll" considers all the musical features. It can be observed that the full set is the one that reaches the highest accuracy rate faster.

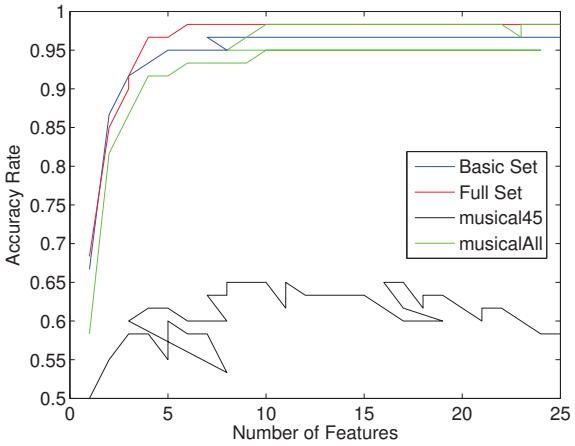


Figure 6.13.: Evaluation of the SES database for the plain Bayes classifier

6.8. Problems and remarks concerning the implementation

Architecture problem In the new architecture separating local and global features, local features have to be computed independently for each block: inter-block operations are not allowed. Therefore, the old implementation for formants and the RAPT pitch estimation algorithm can not be directly re-adapted. The formant set has been removed from the new basic feature set; concerning the pitch, we have used the plain autocorrelation algorithm in the new basic feature and the RAPT algorithm for the musical features, which in theory gives more stable pitch estimations.

Detection of voiced, unvoiced and silent blocks Several features entail computations on voiced, unvoiced and silent blocks. In order to identify which type of block we are dealing with, a combination of the VAD (voice activity detection) algorithm and a voiced detection algorithm is performed.

The voiced detection algorithm is a very simple local feature which compares the energy below and above 1kHz, and takes as voiced all the segments which have more energy in the lower band of the spectrum. Of course, this algorithm alone will not work correctly since all the silent blocks having more energy below 1kHz will be considered as voiced. It gives however good performance when combined with the VAD algorithm.

Missing file in old feature matrix In the old feature matrix, there were only 707 files instead of 708. The file lacking was '10a01Ac.wav' or 'tub_04_m_st_an_mo_lappen_c_16khz.wav' because the pitch estimation algorithm was not able find any pitch at all (no voiced segments are detected because the person in this file is whispering).

Controversy about MFCCs for emotion recognition The usage of MFCCs (defined in Section 4.4) for emotion recognition has met opposition in the literature, as these tend to depend too strongly on the spoken content. This seems a drawback, since we would like to recognize emotions independently of the content. However, MFCCs have proved to be quite accurate in our case; they are even the preferred feature group of our Bayes classifier. These features are better at predicting arousal than valence.

7. Discussion and Conclusions

"Even monkeys express strong feelings in different tones – anger and impatience by low, – fear and pain by high notes." - Charles Darwin

7.1. Summary

This thesis has provided a theoretical and empirical approach to the possible link between emotional speech and music perception. First, an extensive literature review has been provided, along with the description of some of the most relevant experiments in the fields of neuroscience, music theory and linguistics. Musical concepts have been exposed concerning how emotions are transmitted through music.

Secondly, an empirical system for emotion recognition from speech signals has been evaluated. Both traditional and musical features have been implemented. Simulations have proved that musical features can improve the classification accuracy of speech signals. Additionally, a stylization algorithm has been applied to the fundamental frequency F0 in order to get a signal closer to the perceived pitch. The contributions of this thesis can be summarized in the following points:

1. Literature review regarding the relationship between speech, music and emotions.
2. Re-implementation of the basic feature set
3. Implementation of speech processing algorithms
   - Voiced/unvoiced detection
   - Syllabic segmentation
   - Stylization algorithm
4. Validation of the statistical link between music and speech
5. Implementation/improvement of musical features
   - Re-implementation of the interval features and autocorrelation triad features
   - Implementation of the gaussian triad features (selection of dominant tones)
   - Rhythmic features
   - Intensity and timbral features
6. Implementation of a perceptual intonation model (stylization algorithm taking into account perceptual thresholds of the auditory system)

7. Simulations of emotion recognition in speech mostly for German but also for Spanish

7.2. Further research topics

7.2.1. Different environments

Natural emotional speech Natural speech is actually the final target of emotion recognition systems, but it is much harder to analyze than acted speech. First, the range of emotions is unbounded, and the variation degree to express each of them is very important. On the other hand, we might have additional prosodic factors like irregular pauses or affected articulation that might complicate the task of our estimation algorithms like voiced/unvoiced detection or pitch estimation algorithm. Finally, the samples for natural speech are expected to be much noisier than the ones generated in laboratory conditions [58].

Natural speech databases are seldom made publicly available in order to preserve the privacy of the speakers. Creating our own natural speech database would be very expensive both because of the recording and labeling. Yet, there is a database called Naturalistic Belfast Database found during this thesis, already labeled using a specialized software called Praat (see Appendix B). It consists of 239 clips of 10-60s taken from TV or interview recordings in English, from 125 different speakers. This database should prove useful in future research involving natural emotional speech.

Extension of results to other languages or databases It would be interesting to extent our results with musical features to other languages and compare the results. A question that should be investigated is whether musical features present a different behavior for different languages. Our Berlin Emotional Database only comprises 10 different speakers: the range of emotional expression is therefore not very wide. Results on other languages and databases [98][20] would give us information about the degree of universality of the new derived acoustic correlates.

7.2.2. Optimization of the different steps in pattern recognition

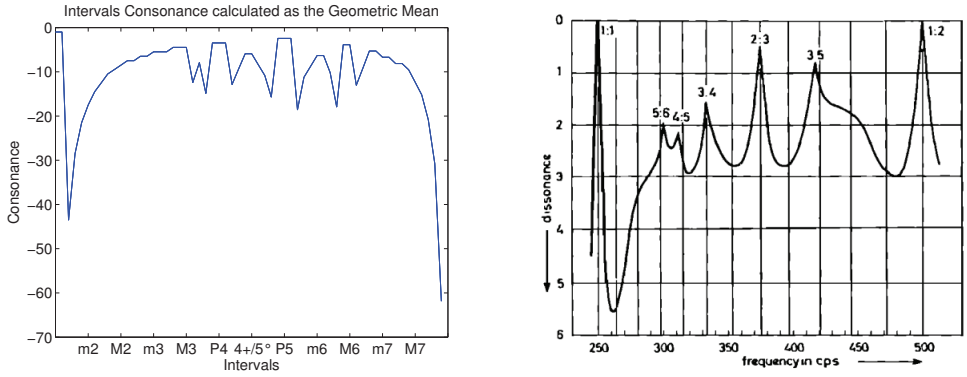
Feature transformation Since the total number of features is very big, they are expected to have an important overlap of information. Our solution so far was to select a subset of features applying the SFFS algorithm. It could nevertheless happen that relevant information scattered across several features be unfortunately discarded. A possible research path could be to transform the whole set of features in order to concentrate the essential information in some principal components. Transforming the feature space might bring interesting properties or inter-feature relationships to light. Some methods like Principal Component Analysis or Linear Discriminant Analysis already implemented in the LSS Toolbox could be used.

Optimization of our speech processing algorithms We have seen that our algorithm RAPT for pitch estimation still presents some errors. Also our implementation to detect voiced, unvoiced and silent blocks is based on a combination of the voice activity detection algorithm and some band energy information. Nowadays, more sophisticated algorithms taking into account the ZCR or the autocorrelation values exist for a more accurate block type classification. Finally, the syllabic segmentation algorithm has not been tested yet: its improvement could give place to more refined musical features.

Classification In our simulations, we observed that some of our features were not suitable for the Bayes classifier: the variance of certain classes is sometimes too small, making the computation of the inverse matrix difficult; other times, distributions present multiple peaks which can not be properly approximated with a single gaussian. It would therefore be interesting to consider other classifiers like Support Vector Machines [82][93][38][77] and observe the behavior of musical features.

7.2.3. Further improvement of the musical features

Optimization of the dissonance model The dissonance model used in our implementation was based on frequency ratios N/D, namely on the geometric mean $\sqrt{N \cdot D}$. This curve is a rough approximation of the empirical curve of Plomp and Levelt. Both curves present many differences, especially on the right side of the graph. We should therefore directly use the empirical curve in order to get more accurate dissonance values.



(a) Approximation curve

(b) Empirical curve

Figure 7.1.: Comparison between perceptual and geometric-mean based consonance

Detection of emotionally meaningful moments in speech Emotional expression is not always present along time. Some temporal moments seems to be more emotionally significant than others. We believe that there might be a strong correlation with loudness or syllable stress. Syllables can be classified in four different categories based on prosodic stress: unstressed, weakly stressed, moderately stressed and heavily stressed. A very

interesting path for future investigation would be the implementation of a syllabic stress detection algorithm, taking into account acoustic characteristics of the signal. A function to detect emotional activity related to energy or stress in syllables could improve the quality of our interval and triad features.

Improvement of the perceptual model of intonation In our simulations, applying stylization to our F0 signal has not improved the quality of interval nor triad features. Figure 7.2 shows that our stylization algorithm is not very robust to pitch estimation errors. Better stylization methods based on linguistics and psychoacoustics should be investigated in order to better approximate the perceived pitch signal. Another path for research would be to use the stylized F0 signal in order to generate features other than the musical ones.

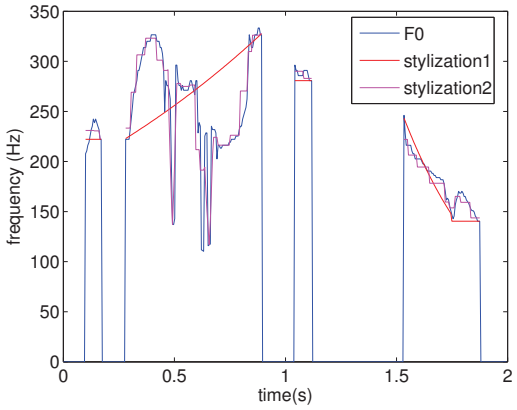


Figure 7.2.: Example of stylization problems

7.2.4. Alternative research paths

Analysis of tonal languages Tonal languages use pitch to signal a difference in meaning between words. Mandarin Chinese is probably the most widely studied tonal language, with five contrasting tones to convey meaning. A direct consequence of that is that short-term F0 changes for emotional expression are less pronounced than in Indo-European languages; emotions seems to be rather expressed through more global tendencies of the pitch.

Recent studies have shown that acoustic correlates of tone at the word level interact with that of the intonation and emotion at the sentence level [17][102]. The analysis of tone languages might therefore give us some information about which aspects of pitch variation are meaningful for affect expression and universals to both speech and music.

Animal vocalization: an evolutionary approach In Section 3.3 on page 28, we have seen how tense triads could resolve into major or minor chords depending on the movement

direction of the middle tone. It should also be noticed that a semitone increase in any tone of an augmented chord results in a minor chord, while a semitone decrease results in a major chord. How can we interpret this observation? How come that the affective valence of major and minor chords transcend cultural boundaries? A possible explanation lies in the so-called *sound symbolism* or *frequency code* [10][11], namely that pitch variations have a certain meaning associated as a result of evolution.

It is believed that music and speech share the same origin in evolution, and some vestiges might be found in the analysis of animal vocalization [25][89][88][32]. Animals tend to signal their strength, aggression or territorial dominance by using vocalizations with a low and/or falling pitch, whereas they convey weakness, defeat and submission using a high and/or rising pitch. Indeed, F0 is inversely related to the mass of the vibrating membrane, which is in turn correlated to the overall body mass: hence big animals normally produce lower pitch calls in comparison with smaller ones.

Analogously, some of these tendencies can also be found in human languages: falling pitch is generally used to signal social strength (commands, assertions, dominance), while increasing pitch often indicates a submissive position (questions, politeness or deference). Indeed, *"ascending contours convey uncertainty and uneasiness, and descending contours certainty and stability"* [10].

Can this hypothesis be verified? Is this sound symbolism still present in emotional speech? Analysis of chimpanzee's calls for instance and comparison with properties of the human speech could bring some insight on this issue. Generally, animals can produce more or less harmonic vocalizations [94][26]. Contact or affective barks are generally tonal and harmonically rich, whereas alarm barks tend to be noisy and harsh. There is a gradation in the signaler's emotion depending on the periodicity and harmony of the calls.

	Animal vocalizations	Human languages	Musical harmony
Rising pitch	weakness defeat submission	politeness assent questions	negative affect, despair, sadness
Falling pitch	dominance strength victory	commands assertions statements	positive affect, joy, happiness

Table 7.1.: Sound symbolism in animal calls, speech and music

Analysis of emotional transmission through music Given a certain database with small music extracts and simple labels (happy, angry, scared, sad), it would be interesting to look for features that best correlate to emotions in music. During this thesis, such a database was never found, but both fields are getting closer and closer, and it might be easier to find it in the future. There are also some efforts to generate a uniform taxonomy for musical emotions [42], as can be seen in Table 7.2 on the next page.

Clusters	Mood Adjectives
1	passionate, rousing, confident, boisterous, rowdy
2	rollicking, cheerful, fun, sweet, amiable/good natured
3	literate, poignant, wistful, bittersweet, autumnal, brooding
4	humorous, silly, campy, quirky, whimsical, witty, wry
5	aggressive, fiery, tense/anxious, intense, volatile, visceral

Table 7.2.: Example of a classification of musical emotions into 5 clusters

Genetic Algorithms Our final suggestion for the future is the implementation of an automatic way to generate several features [75]. Feature mining can be a troublesome and time-consuming task. We might never find good features if we do not automatize the search. Genetic algorithms are therefore desirable, in that they provide flexibility and systematization in the process.

Appendix

A. Glossary related to musical features

Fundamental Frequency F0 Inverse of the smallest true period in the interval being analyzed.

Pitch Auditory percept of tone, usually approximated by the fundamental frequency F0.

Prosody Rhythm, stress, and intonation of speech. the patterns of stress and intonation in a language, Melody of speech determined primarily by modifications of pitch, quality, strength, and duration; perceived primarily as stress and intonational patterns. Encoding for emotional state, nature of utterance (statement, question, command), irony/sarcasm, emphasis/contrast/focus.

Critical Bandwidth Bandwidth of each of our auditory filters; information collection and integration unit on the basilar membrane, corresponding to a constant number of 1300 receptor cells. It corresponds to the frequency distance above which the roughness sensation between two pure tones disappears.

Tonality Hierarchical ordering of the pitches of the chromatic scale such that these notes are perceived in relation to one central and stable pitch called the tonic.

Standardized Key Profile Patterns of perceived stability of pitches.

Triad Combination of three-tone in a chord (simultaneously played).

Dissonance Measure of the degree of pleasantness caused by an interval or a chord.

Tension In music, tension is the perceived need for relaxation or release created by a listener's expectations. For example, dissonance may give way to consonance. Tension may also be produced through reiteration or gradual motion to a higher pitch.

Modality Measure of valence degree of a triad.

Rhythm Systematic patterns of timing, accent, grouping in speech.

B. Software available for audio signal processing

- Praat: tool for phonetics research (http://www.fon.hum.uva.nl/praat/)
- Colea: Matlab software tool for speech analysis
  (http://www.utdallas.edu/~loizou/speech/colea.htm)
  Interesting for the computation of formants and F0.
- Marsyas: Musical Research System for Analysis and Synthesis
- Mir toolbox: A toolbox for musical feature extraction from audio
  https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox/mirtoolbox
- Psysound: software for the analysis of sound recordings using physical and psychoacoustical algorithms

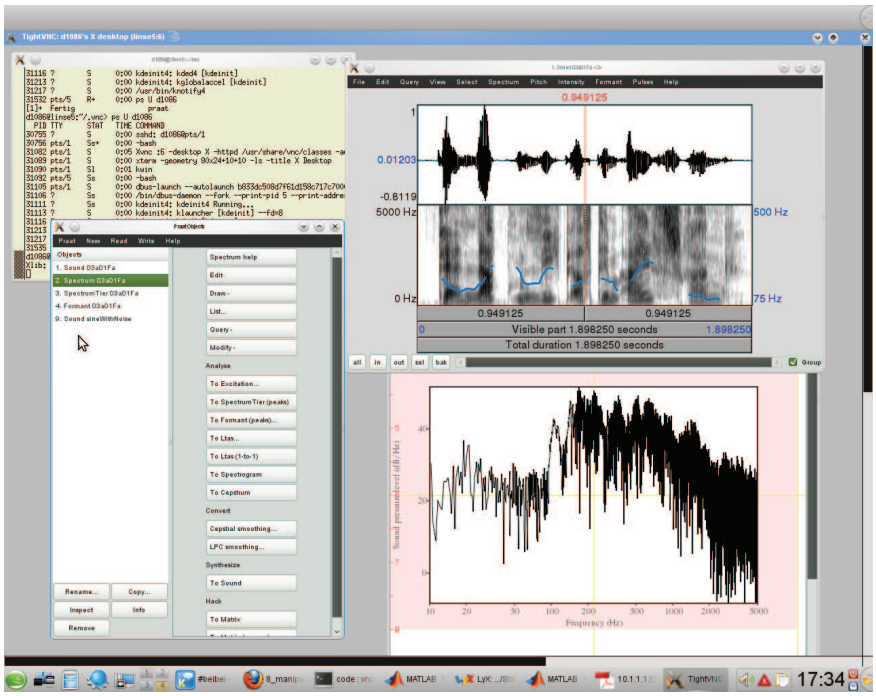


Figure B.1.: Praat software

C. Other results

Histogram matching approach

For each emotion, we compute the global circular pitch histogram. Each emotion is thus represented with an emotional pattern P_i. For each utterance in the test set, let us compute the individual circular pitch histogram, and try to deduce from which distribution it comes from.

Solution for this statistical problem (maximum likelihood principle): take the emotion i which maximizes the probability of having the individual observation/histogram x given the emotional pattern P_i.

$$emotion = argmax_i \, p\,(x|P_i)$$

In this experiment, we observe that the emotional patterns are very similar between classes, which is not very promising. We believe that the reasons are the following:

1. Prosodic interferences: prosody in speech can be seen as a single channel encoding several functions others than emotion. As it was explained in Section 2.3 on page 16, even neutral speech has a rich histogram of melodic intervals. This can be understood as noise or interferences on the emotional channel.

2. Ambiguous clusters: so far, we are working with general classes like angry or sad, but there is actually multiple ways to express the same emotion. Anger for instance can be cold or hot. This is a problem in our initial model of emotions, difficult to overcome. Apparently, the intra-class variance is bigger than the inter-class variance. In this perspective, the emotional model should be reviewed.

Principal pitch interval

Emotions are not always constantly expressed, precise moments seem to concentrate more emotional information than others. Therefore, we would like to distinguish between *purely semantic* speech and *emotionally meaningful* speech. With this premise in mind, we should ask ourselves about the nature of these meaningful moments: are they correlated with the loudness, energy or the syllabic stress? Our hypothesis is that some intervals have more meaning than others. Here we tested the influence of the pitch interval at maximum energy:

$$princInt = pitch_{MaxEn} - \text{mean}\,(pitch)$$

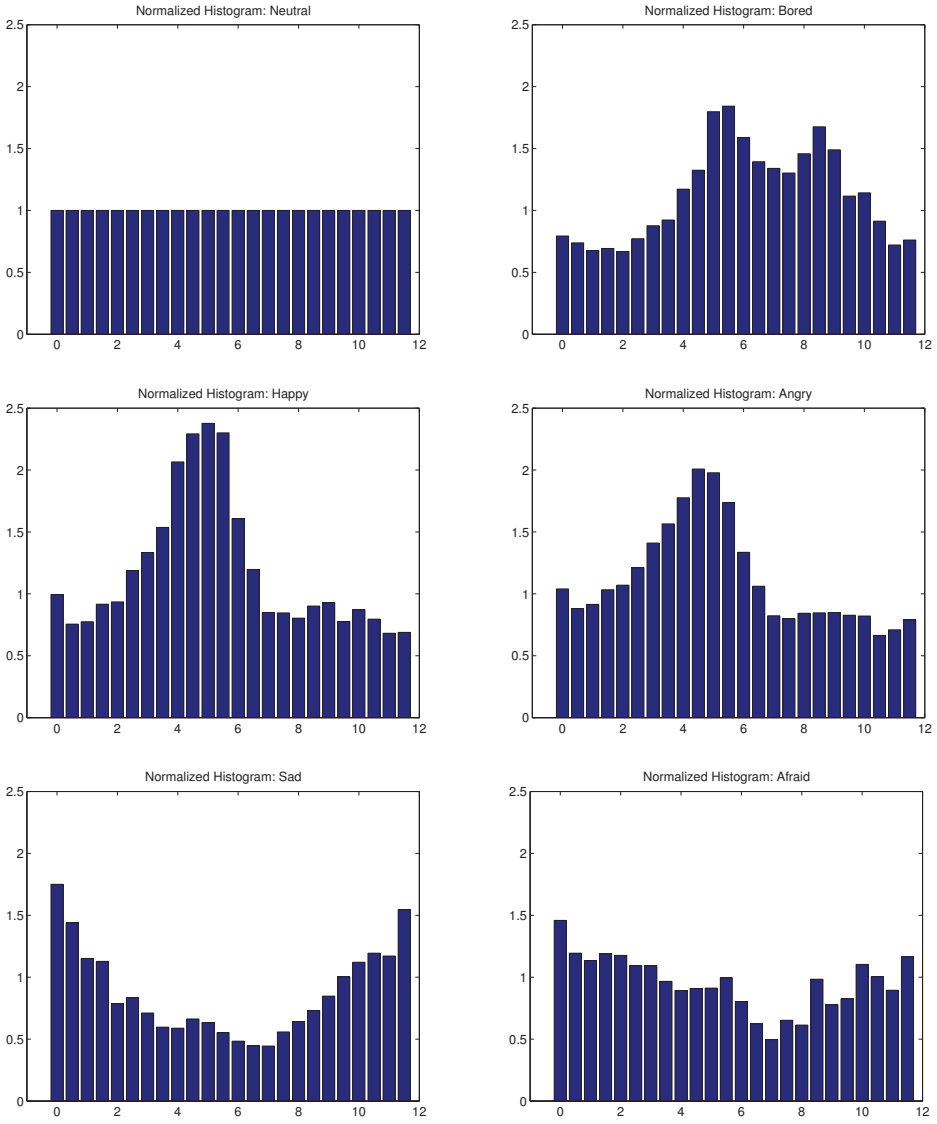


Figure C.1.: Normalized emotional patterns

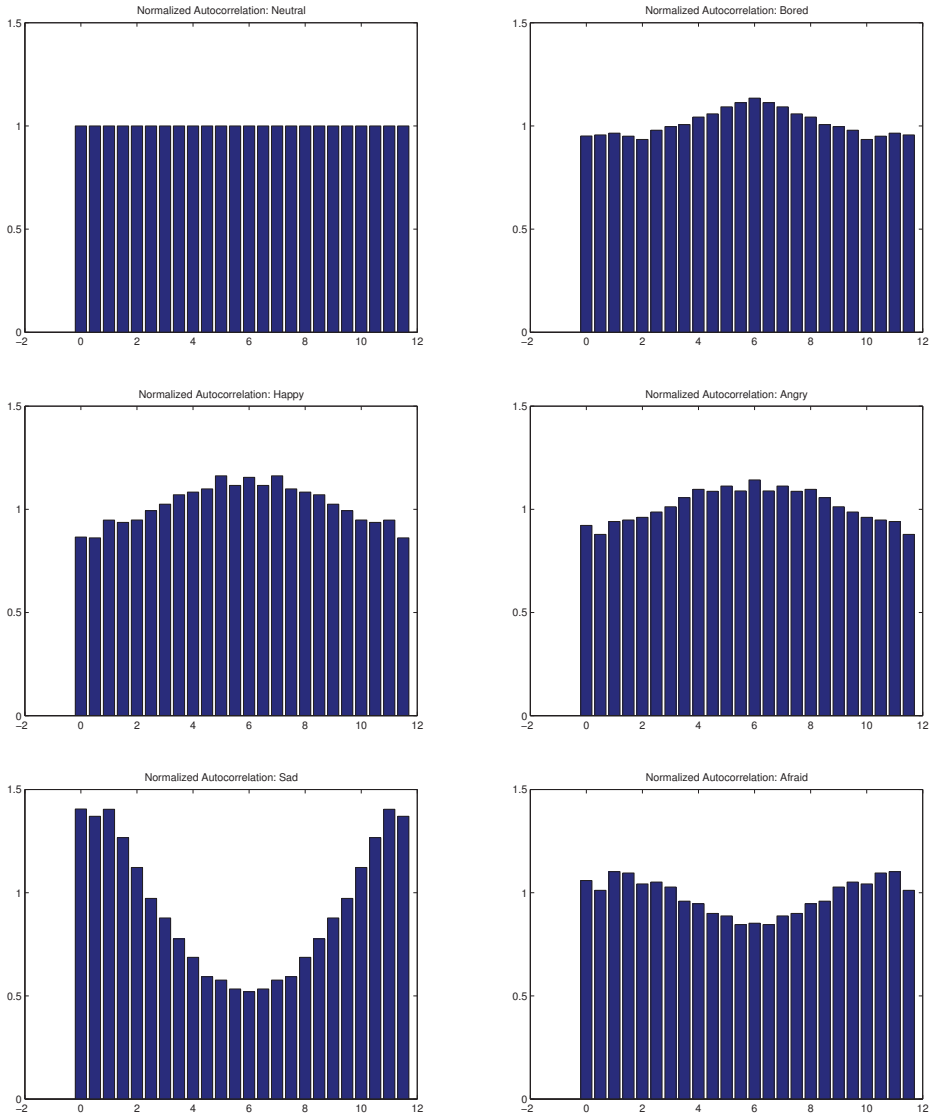
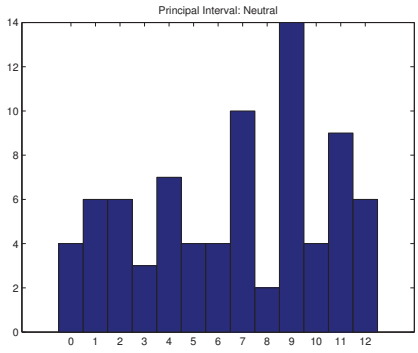
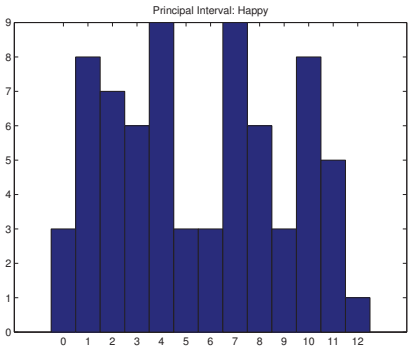


Figure C.2.: Normalized distribution of intervals





Conclusion: it does not seem to work; the histograms are not representative.

Temporal dynamics for pitch

Mean transition probability matrix

State k \ k+1	Rise	Fall	Max	Min
Rising Slope	0	0.343	0.0398	0
Falling Slope	0.2277	0	0.1555	0.0036
Max Plateau	0.1539	0.0438	0	0.0317
Min Plateau	0.0005	0	0.0005	0

Rise	Fall	Max	Min
0.3701	0.3740	0.2218	0.0341

It can be observed that all emotional classes have very similar probabilities. We believe that the variation between sentences is too big. Therefore we performed this analysis for each sentence separately.

Transition probability matrix for one single sentence

Sentence: “Das will Sie am Mittwoch abgeben”.

Neutral	Rise	Fall	Max	Min
Rise	0	58	6	0
Fall	35	0	31	0
Max	29	8	0	9
Min	0	0	0	0

Happy	Rise	Fall	Max	Min
Rise	0	38	3	0
Fall	18	0	22	0
Max	23	2	0	6
Min	0	0	0	0

	Rise	Fall	Max	Min
Happy	27.40	37.19	34.27	1.13
Angry	23.05	38.77	37.09	1.09

Table C.1.: Duration probabilities for the sentence “Das will sie am Mittwoch abgeben”

Happy	Rise	Fall	Max
Rise	0	95.24	4.76
Fall	47.62	0	52.38
Max	73.33	6.67	0

Angry	Rise	Fall	Max
Rise	0	96.81	3.19
Fall	46.88	0	52.08
Max	73.13	7.46	0

Table C.2.: Event probabilities for the sentence “Das will sie am Mittwoch abgeben”

Sad	Rise	Fall	Max	Min
Rise	0	40	7	0
Fall	23	0	23	2
Max	23	8	0	6
Min	1	0	0	0

Bored	Rise	Fall	Max	Min
Rise	0	47	6	0
Fall	30	0	21	3
Max	23	7	0	6
Min	0	0	1	0

Angry	Rise	Fall	Max	Min
Rise	0	93	4	0
Fall	46	0	50	0
Max	51	3	0	14
Min	0	0	0	0

Afraid	Rise	Fall	Max	Min
Rise	0	37	2	0
Fall	20	0	19	0
Max	19	2	0	7
Min	0	0	0	0

	Rise	Fall	Max	Min
Neutral	64	66	46	9
Happy	41	40	31	6
Sad	47	48	37	8
Bored	53	54	36	9
Angry	97	96	68	11
Afraid	39	39	28	7

This research path does not look very promising since the inter-class distances are not very significant. For this experiment, we used the small emoDB, which has 10 files for each emotion in the case of the sentence “Das will Sie am Mittwoch abgeben”.

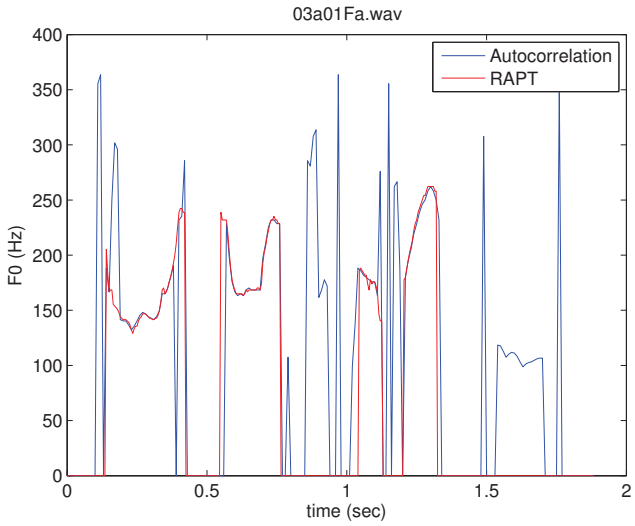


Figure C.3.: Comparison of the two pitch estimation algorithms

Comparison of the available pitch estimation methods

Plain autocorrelation method Vs RAPT

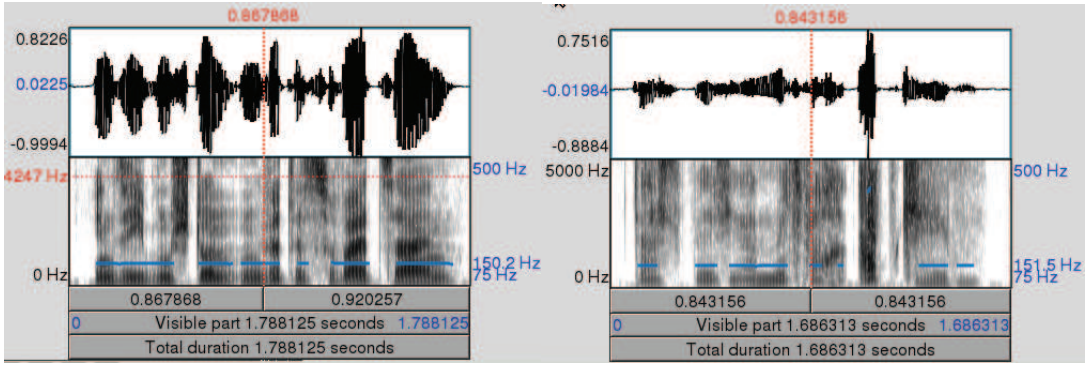
In our implementation, with use both methods: the plain auto-correlation is used for the computation of basic pitch-based features whereas RAPT is used for musical features. The plain autocorrelation is easier to compute and works under the new architecture (local and global features). Nevertheless, it is less exact. The RAPT (Robust Algorithm for Pitch Tracking) uses a two-level auto-correlation and dynamic programming to find the best path. It is therefore more robust and suitable for music-theory-based computations. We can see that the auto-correlation method gives similar results for the overlapping areas.

RAPT Vs PRAAT algorithm

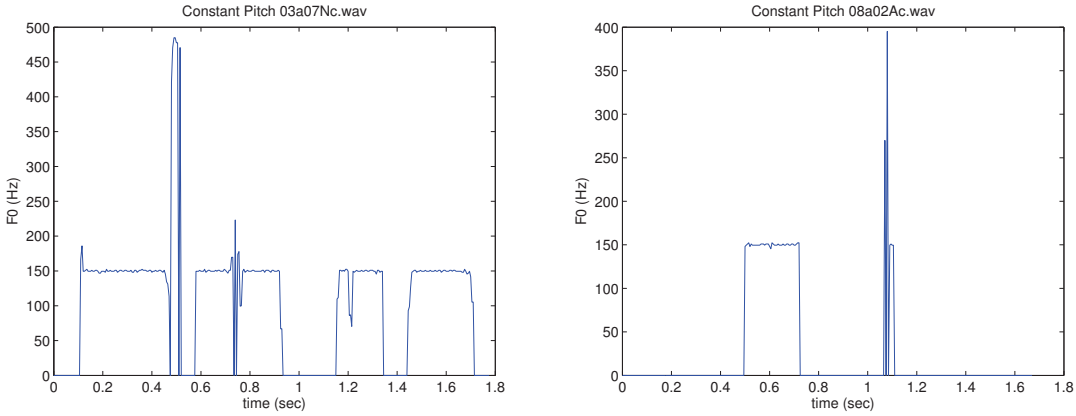
Here we compare the pitch estimation for the RAPT algorithm and the one used in the software Praat. Both methods are based on the autocorrelation and dynamic programming. The following plots show a constant pitch for two different audio files. We have indeed modified the fundamental frequency F0 signal artificially (using the software Praat) and then extracted the F0 signal using both the RAPT and the algorithm of Praat.

It can be observed that our current algorithm RAPT is less accurate than the one of Praat. Indeed, we get two different types of errors:

- Presence of high peaks corresponding to unvoiced segments
- No detection of certain segments



(a) Pitch estimation using the algorithm contained in the Praat software



(b) Pitch estimation using the RAPT algorithm

Figure C.4.: Comparison of pitch estimation algorithms

List of available musical features

```
  1 - DurVMean
  2 - DurAMean
  3 - DurSMean
  4 - DurVStd
  5 - DurAStd
  6 - DurSStd
  7 - DurVTot2DurTot
  8 - DurATot2DurTot
  9 - DurSTot2DurTot
 10 - DurVSpeed
 11 - DurASpeed
 12 - DurSSpeed
 13 - DurVMean2SMean
 14 - DurAMean2SMean
 15 - DurVLen2ALen
 16 - DurVLen2SLen
--------------------
 17 - MFCC1_mean
 18 - MFCC1_std
 19 - MFCC1_diffMean
 20 - MFCC1_diffStd
 21 - MFCC1_diff2Mean
 22 - MFCC1_diff2Max
 23 - MFCC1_diff2Std
 24 - MFCC2_mean
 25 - MFCC2_std
 26 - MFCC2_diffMean
 27 - MFCC2_diffStd
 28 - MFCC2_diff2Mean
 29 - MFCC2_diff2Max
 30 - MFCC2_diff2Std
 31 - MFCC3_mean
 32 - MFCC3_std
 33 - MFCC3_diffMean
 34 - MFCC3_diffStd
 35 - MFCC3_diff2Mean
 36 - MFCC3_diff2Max
 37 - MFCC3_diff2Std
 38 - MFCC4_mean
 39 - MFCC4_std
 40 - MFCC4_diffMean
 41 - MFCC4_diffStd
 42 - MFCC4_diff2Mean
 43 - MFCC4_diff2Max
 44 - MFCC4_diff2Std
 45 - MFCC5_mean
 46 - MFCC5_std
 47 - MFCC5_diffMean
 48 - MFCC5_diffStd
 49 - MFCC5_diff2Mean
 50 - MFCC5_diff2Max
 51 - MFCC5_diff2Std
 52 - MFCC6_mean
 53 - MFCC6_std
 54 - MFCC6_diffMean
 55 - MFCC6_diffStd
 56 - MFCC6_diff2Mean
 57 - MFCC6_diff2Max
 58 - MFCC6_diff2Std
 59 - MFCC7_mean
 60 - MFCC7_std
 61 - MFCC7_diffMean
 62 - MFCC7_diffStd
 63 - MFCC7_diff2Mean
 64 - MFCC7_diff2Max
 65 - MFCC7_diff2Std
 66 - MFCC8_mean
 67 - MFCC8_std
 68 - MFCC8_diffMean
 69 - MFCC8_diffStd
 70 - MFCC8_diff2Mean
 71 - MFCC8_diff2Max
 72 - MFCC8_diff2Std
 73 - MFCC9_mean
 74 - MFCC9_std
 75 - MFCC9_diffMean
 76 - MFCC9_diffStd
 77 - MFCC9_diff2Mean
 78 - MFCC9_diff2Max
 79 - MFCC9_diff2Std
 80 - MFCC10_mean
 81 - MFCC10_std
 82 - MFCC10_diffMean
 83 - MFCC10_diffStd
 84 - MFCC10_diff2Mean
 85 - MFCC10_diff2Max
 86 - MFCC10_diff2Std
 87 - MFCC11_mean
 88 - MFCC11_std
 89 - MFCC11_diffMean
 90 - MFCC11_diffStd
 91 - MFCC11_diff2Mean
 92 - MFCC11_diff2Max
 93 - MFCC11_diff2Std
 94 - MFCC12_mean
 95 - MFCC12_std
 96 - MFCC12_diffMean
 97 - MFCC12_diffStd
 98 - MFCC12_diff2Mean
 99 - MFCC12_diff2Max
```

100 - MFCC12_diff2Std
101 - MFCC13_mean
102 - MFCC13_std
103 - MFCC13_diffMean
104 - MFCC13_diffStd
105 - MFCC13_diff2Mean
106 - MFCC13_diff2Max
107 - MFCC13_diff2Std
---
108 - mzcrMean
109 - mzcrMax
110 - mzcrMedian
111 - mzcrStd
112 - mzcrIqr
113 - mzcrDiffMin
114 - mzcrDiffMax
115 - mzcrDiffStd
116 - mzcrDiffIqr
117 - mzcrDiff2Min
118 - mzcrDiff2Max
119 - mzcrDiff2Std
120 - mzcrDiff2Iqr
---
121 - vHarmonyMean
122 - vHarmonyStd
123 - vHarmonyIqr
---
124 - mEnMean
125 - mEnMax
126 - mEnMedian
127 - mEnStd
128 - mEnIqr
129 - mEnDiffMin
130 - mEnDiffMax
131 - mEnDiffStd
132 - mEnDiffIqr
133 - mEnDiff2Min
134 - mEnDiff2Max
135 - mEnDiff2Std
136 - mEnDiff2Iqr
137 - vEnMax
138 - vEnMean
139 - vEnMedian
140 - vEnStd
141 - vEnIqr
142 - vEnDiffMin
143 - vEnDiffMax
144 - vEnDiffStd
145 - vEnDiffIqr
146 - vEnDiff2Min
147 - vEnDiff2Max
148 - vEnDiff2Std
149 - vEnDiff2Iqr
150 - aEnMean
151 - vEnFirstMean
152 - vEnLastMean
153 - EnVInt2SInt
154 - EnAInt2SInt
155 - mEnBelow250Mean
156 - mEnBand1/5Mean
157 - mEnBand2/5Mean
158 - mEnBand3/5Mean
159 - mEnBand4/5Mean
160 - mEnBand5/5Mean
161 - mEnBelow250Std
162 - mEnBand1/5Std
163 - mEnBand2/5Std
164 - mEnBand3/5Std
165 - mEnBand4/5Std
166 - mEnBand5/5Std
167 - aEnPlatMaxMean
168 - aEnSlopeRaiseMean
169 - aEnSlopeRaiseMedian
170 - aEnSlopeRaiseIqr
171 - aEnSlopeFallMean
172 - aEnSlopeFallMedian
173 - aEnSlopeFallIqr
174 - aEnDurSlopeRaiseMax
175 - aEnDurSlopeRaiseMean
176 - aEnDurSlopeRaiseMedian
177 - aEnDurSlopeRaiseIqr
178 - aEnDurSlopeFallMax
179 - aEnDurSlopeFallMean
180 - aEnDurSlopeFallMedian
181 - aEnDurSlopeFallIqr
---
182 - vPitchMin
183 - vPitchMax
184 - vPitchMean
185 - vPitchMedian
186 - vPitchStd
187 - vPitchIqr
188 - vPitchDiffMin
189 - vPitchDiffMax
190 - vPitchDiffMean
191 - vPitchDiffStd
192 - vPitchDiffIqr
193 - vPitchDiff2Min
194 - vPitchDiff2Max
195 - vPitchDiff2Mean

196 - vPitchDiff2Std
197 - vPitchDiff2Iqr
198 - vPitchFirstMean
199 - vPitchLastMean
200 - mPitchPlatMaxMean
201 - mPitchSlopeRaiseMean
202 - mPitchSlopeRaiseMedian
203 - mPitchSlopeRaiseIqr
204 - mPitchSlopeFallMean
205 - mPitchSlopeFallMedian
206 - mPitchSlopeFallIqr
207 - mPitchDurSlopeRaiseMax
208 - mPitchDurSlopeRaiseMean
209 - mPitchDurSlopeRaiseMedian
210 - mPitchDurSlopeRaiseIqr
211 - mPitchDurSlopeFallMax
212 - mPitchDurSlopeFallMean
213 - mPitchDurSlopeFallMedian
214 - mPitchDurSlopeFallIqr
---
215 - CPDC1
216 - CPDC2
217 - CPDC3
218 - CPDC4
219 - CPDC5
220 - CPDC6
221 - CPDC7
222 - CPDC8
223 - CPDC9
224 - CPDC10
225 - CPDC11
226 - CPDC12
227 - CPDC13
228 - CPDC14
229 - CPDC15
230 - CPDC16
231 - CPDC17
232 - CPDC18
233 - CPDC19
234 - CPDC20
235 - CPDC21
236 - CPDC22
237 - CPDC23
238 - CPDC24
239 - CPDC25
240 - CPDC26
241 - CPDC27
242 - CPDC28
243 - CPDC29
244 - CPDC30
245 - totalDis
---
246 - dur
247 - moll
248 - tension
249 - modality
---
250 - cookDis
251 - cookTension
252 - cookModality
253 - aTon
254 - aMin_Ton
255 - aMed_Ton
256 - sigmaMin_Ton
257 - sigmaMed_Ton
258 - FR1
259 - FR2
---
260 - IR1_mean
261 - IR1_std
262 - IR1_diffMean
263 - IR1_diffStd
264 - IR1_diff2Mean
265 - IR1_diff2Max
266 - IR1_diff2Std
267 - IR2_mean
268 - IR2_std
269 - IR2_diffMean
270 - IR2_diffStd
271 - IR2_diff2Mean
272 - IR2_diff2Max
273 - IR2_diff2Std
274 - IR3_mean
275 - IR3_std
276 - IR3_diffMean
277 - IR3_diffStd
278 - IR3_diff2Mean
279 - IR3_diff2Max
280 - IR3_diff2Std
281 - IR4_mean
282 - IR4_std
283 - IR4_diffMean
284 - IR4_diffStd
285 - IR4_diff2Mean
286 - IR4_diff2Max
287 - IR4_diff2Std
288 - IR5_mean
289 - IR5_std
290 - IR5_diffMean
291 - IR5_diffStd

292 - IR5_diff2Mean
293 - IR5_diff2Max
294 - IR5_diff2Std
295 - IR6_mean
296 - IR6_std
297 - IR6_diffMean
298 - IR6_diffStd
299 - IR6_diff2Mean
300 - IR6_diff2Max
301 - IR6_diff2Std
302 - IR7_mean
303 - IR7_std
304 - IR7_diffMean
305 - IR7_diffStd
306 - IR7_diff2Mean
307 - IR7_diff2Max
308 - IR7_diff2Std
309 - IR8_mean
310 - IR8_std
311 - IR8_diffMean
312 - IR8_diffStd
313 - IR8_diff2Mean
314 - IR8_diff2Max
315 - IR8_diff2Std
316 - I_mean
317 - I_std
318 - I_diffMean
319 - I_diffStd
320 - I_diff2Mean
321 - I_diff2Max
322 - I_diff2Std
---
323 - rhythmStrMin
324 - rhythmStrMean
325 - rhythmStrMax
326 - rhythmStrMedian
327 - rhythmStrStd
328 - rhythmRegMin
329 - rhythmRegMean
330 - rhythmRegMax
331 - rhythmRegMedian
332 - rhythmRegStd
333 - rhythmSpeMin
334 - rhythmSpeMean
335 - rhythmSpeMax
336 - rhythmSpeMedian
337 - rhythmSpeStd
---
338 - sF0_CPDC1
339 - sF0_CPDC2
340 - sF0_CPDC3
341 - sF0_CPDC4
342 - sF0_CPDC5
343 - sF0_CPDC6
344 - sF0_CPDC7
345 - sF0_CPDC8
346 - sF0_CPDC9
347 - sF0_CPDC10
348 - sF0_CPDC11
349 - sF0_CPDC12
350 - sF0_CPDC13
351 - sF0_CPDC14
352 - sF0_CPDC15
353 - sF0_CPDC16
354 - sF0_CPDC17
355 - sF0_CPDC18
356 - sF0_CPDC19
357 - sF0_CPDC20
358 - sF0_CPDC21
359 - sF0_CPDC22
360 - sF0_CPDC23
361 - sF0_CPDC24
362 - sF0_CPDC25
363 - sF0_CPDC26
364 - sF0_CPDC27
365 - sF0_CPDC28
366 - sF0_CPDC29
367 - sF0_CPDC30
368 - sF0_totalDis
369 - sF0_dur
370 - sF0_moll
371 - sF0_tension
372 - sF0_modality
373 - sF0_cookDis
374 - sF0_cookTension
375 - sF0_cookModality
376 - sF0_aTon
377 - sF0_aMin_Ton
378 - sF0_aMed_Ton
379 - sF0_sigmaMin_Ton
380 - sF0_sigmaMed_Ton
381 - sF0_FR1
382 - sF0_FR2

Bibliography

[1] Chroma-based estimation of musical key from audio-signal analysis.

[2] Jo-Anne Bachorowski and Michael J. Owren. Sounds of Emotion. *Annals of the New York Academy of Sciences*, 1000(1):244–265, January 2006.

[3] R Banse and K R Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614–36, March 1996.

[4] Charles A Bouman. Cluster: An Unsupervised Algorithm for Modeling Gaussian Mixtures. pages 1–20, 2005.

[5] Steven Brown, Michael J Martinez, and Lawrence M Parsons. Music and language side by side in the brain: a PET study of the generation of melodies and sentences. *The European journal of neuroscience*, 23(10):2791–803, May 2006.

[6] Murtaza Bulut and Shrikanth Narayanan. On the robustness of overall F0-only modifications to the perception of emotions in speech. *The Journal of the Acoustical Society of America*, 123(6):4547–58, June 2008.

[7] Alain De Cheveign. Comparative evaluation of F0 estimation algorithms. *Methods*, 2001.

[8] Elaine Chew. Modeling tonality: applications to music cognition. *Systems Engineering*, pages 4–9.

[9] N.D. Cook. *Tone of voice and mind: the connections between intonation, emotion, cognition, and consciousness*. John Benjamins Publishing Company, 2002.

[10] Norman D Cook. The psychophysics of harmony perception: Harmony is a three-tone phenomenon. 1(2):106–126, 2006.

[11] Norman D. Cook. The sound symbolism of major and minor harmonies. *New York*, pages 315–319, 2007.

[12] Norman D Cook, Takashi Fujisawa, and Kazuaki Takami. A Psychophysical Model of Harmony Perception. *Main*, pages 493–496, 2004.

[13] Norman D Cook and Takashi X Fujisawa. The Use of Multi-pitch Patterns for Evaluating the Positive and Negative Valence of Emotional Speech. *Work*, (Figure 2):5–8, 2006.

[14] R. Cowie and E. Douglas-Cowie. Automatic statistical analysis of the signal and prosodic signs of emotion in speech. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, pages 1989–1992.

[15] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, K. Kollias, W. Fellenz, and J.G. Taylor. Emotion recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, 22(1), January 2005.

[16] C D'Alessandro. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech & Language*, 9(3):257–288, July 1995.

[17] Chinar Dara and Marc D Pell. The interaction of linguistic and affective prosody in a tone language. *The Journal of the Acoustical Society of America*, 119(5):3303–3304, 2006.

[18] Nicola Dibben. Emotion and music: A view from the cultural psychology of music. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–3, September 2009.

[19] Ellen Dissanayake. A review of The singing neanderthals: The origins of music, language, mind and body. *Evolutionary Psychology*, pages 375–380, 2005.

[20] E Douglas-Cowie. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):33–60, April 2003.

[21] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992.

[22] Evelina Fedorenko, Aniruddh Patel, Daniel Casasanto, Jonathan Winawer, and Edward Gibson. Structural integration in language and music: evidence for a shared system. *Memory & cognition*, 37(1):1–9, January 2009.

[23] Steven Feld and Aaron a. Fox. Music and Language. *Annual Review of Anthropology*, 23(1):25–53, October 1994.

[24] Raul Fernandez. *A computational model for the automatic recognition of affect in speech*. PhD thesis, Massachusetts Institute of Technology, 2004.

[25] Claudia Fichtel, Kurt Hammerschmidt, and Uwe Jürgens. on the Vocal Expression of Emotion. a Multi-Parametric Analysis of Different States of Aversion in the Squirrel Monkey. *Behaviour*, 138(1):97–116, January 2001.

[26] W Fitch. Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour*, 63(3):407–418, March 2002.

[27] Fabian Friedrichs. Schaetzung von prosodischen Features zur Emotionsdetektion Gliederung.

[28] Thomas Fritz, Sebastian Jentschke, Nathalie Gosselin, Daniela Sammler, Isabelle Peretz, Robert Turner, Angela D Friederici, and Stefan Koelsch. Universal recognition of three basic emotions in music. *Current biology : CB*, 19(7):573–6, April 2009.

[29] S. Fukuda and V. Kostov. Extracting emotion from voice. *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028)*, pages 299–304.

[30] Bachu R G, Kopparthi S, Adapa B, and Barkana B D. Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal. pages 1–7.

[31] David Gerhard. Pitch Extraction and Fundamental Frequency : History and Current Techniques Theory of Pitch. *Time*, pages 0–22, 2003.

[32] Aa Ghazanfar and Md Hauser. The neuroethology of primate vocal communication: substrates for the evolution of speech. *Trends in Cognitive Sciences*, 3(10):377–384, 1999.

[33] Oliver Grewe and Frederik Nagel. Individual emotional reactions towards music: Evolutionary-based universals ? *Society*, (2009):261–287, 2010.

[34] Oliver Grewe, Frederik Nagel, Reinhard Kopiez, and Eckart Altenmüller. How does music arouse "chills"? Investigating strong emotions, combining psychological, physiological, and psychoacoustical methods, December 2005.

[35] J. T. Hart, R. Collier, and a. Cohen. A perceptual study of intonation, 1990.

[36] C Hoelper, A Frankort, and C Erdmann. Voiced, unvoiced, silence classification for offline speech coding. *Signal Processing*, pages 4–5.

[37] Patrick G. Hunter, E. Glenn Schellenberg, and Ulrich Schimmack. Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions. *Psychology of Aesthetics, Creativity, and the Arts*, 4(1):47–56, 2010.

[38] Theodoros Iliou and Christos-Nikolaos Anagnostopoulos. Comparison of Different Classifiers for Emotion Recognition. *2009 13th Panhellenic Conference on Informatics*, pages 102–106, 2009.

[39] John Jay, Hopkins Drive, and San Diego. Language and Music as Cognitive Systems. *Language*, pages 1–41, 2009.

[40] R Joseph. The right cerebral hemisphere: emotion, music, visual-spatial skills, body-image, dreams, and awareness. *Journal of clinical psychology*, 44(5):630–73, September 1988.

[41] Patrik N Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychological bulletin*, 129(5):770–814, September 2003.

[42] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music Emotion Recognition: a state of the art review. *Information Retrieval*, (Ismir):255–266, 2010.

[43] W. Baastian Kleijn, Kuldip K. Paliwal, and Talkin David. *A Robust Algorithm for Pitch Tracking*, chapter 14, pages 497–516. Elsevier Science B.V., 1995.

[44] Stefan Koelsch, Katrin Schulze, Daniela Sammler, Thomas Fritz, Karsten Müller, and Oliver Gruber. Functional architecture of verbal and tonal working memory: an FMRI study. *Human brain mapping*, 30(3):859–73, March 2009.

[45] Mark D Korhonen, David a Clausi, and M Ed Jernigan. Modeling emotional content of music using system identification. *IEEE transactions on systems, man, and cybernetics. IEEE Systems, Man, and Cybernetics Society*, 36(3):588–99, June 2006.

[46] C L Krumhansl. Music psychology:tonal structures in perception and memory. *Annual review of psychology*, 42:277–303, January 1991.

[47] Fred Lerdahl and Carol L Krumhansl. Modeling tonal tension. pages 329–366, 2007.

[48] Michael Lewis, J.M. Haviland-Jones, and L.F. Barrett. *Handbook of emotions*. The Guilford Press, 2008.

[49] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03*, page 282, 2003.

[50] D. Liu. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):5–18, January 2006.

[51] Dan Liu. Automatic Mood Detection from Acoustic Music Data 1. *Stress: The International Journal on the Biology of Stress*, 2003.

[52] Steven Robert Livingstone and William Forde. The emergence of music from the Theory of Mind. *Musicae Scientiae*, (2009):83–115, 2010.

[53] Marko Lugger. *Mehrstufige Klassifikation paralinguistischer Eigenschaften aus Sprachsignalen mit Hilfe neuartiger Merkmale.* PhD thesis, Stuttgart, 2010.

[54] P Mertens. Pitch contour stylization using a tonal perception model. *Phonetica.*

[55] D Morrison, R Wang, and L Desilva. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112, February 2007.

[56] S J L Mozziconacci and D J Hermes. A study of intonation patterns in speech expressing emotion or attitude : Production and perception. *Hermes*, pages 154–160, 1997.

[57] Frederik Nagel, Reinhard Kopiez, Oliver Grewe, and Eckart Altenmueller. Psychoacoustical correlates of musically induced chills. *European Society for the Cognitive Sciences of Music*, XII:101–113, 2008.

[58] Daniel Neiberg, Kjell Elenius, Inger Karlsson, and Kornel Laskowski. Emotion Recognition in Spontaneous Speech. *Computer*, 52:101–104, 2006.

[59] Francis Nolan. Intonational equivalence : an experimental evaluation of pitch scales. pages 771–774, 2003.

[60] M. Ogihara. Content-based music similarity search and emotion detection. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages V–705–8, 2004.

[61] A Paeschke, M Kienast, W F Sendlmeier, and Technische Universität Berlin. F0 contours in emotional speech.

[62] Astrid Paeschke. Global Trend of Fundamental Frequency in Emotional Speech. *Science*, pages 6–9, 2004.

[63] a D Patel, E Gibson, J Ratner, M Besson, and P J Holcomb. Processing syntactic relations in language and music: an event-related potential study. *Journal of cognitive neuroscience*, 10(6):717–33, November 1998.

[64] a D Patel, I Peretz, M Tramo, and R Labreque. Processing prosodic and musical patterns: a neuropsychological investigation. *Brain and language*, 61(1):123–44, January 1998.

[65] A.D. Patel. *Music, language, and the brain.* Oxford Univ Pr, 2008.

[66] Aniruddh D Patel. Language, music, syntax and the brain. *Nature neuroscience*, 6(7):674–81, July 2003.

[67] Aniruddh D Patel and Joseph R Daniele. An empirical comparison of rhythm in language and music. *Neurosciences*, 87, 2003.

[68] Aniruddh D Patel, John R Iversen, Micah R Bregman, and Irena Schulz. Studying synchronization to a musical beat in nonhuman animals. *Annals of the New York Academy of Sciences*, 1169:459–69, July 2009.

[69] Aniruddh D. Patel, John R. Iversen, and Jason C. Rosenberg. Comparing the rhythm and melody of speech and music: The case of British English and French. *The Journal of the Acoustical Society of America*, 119(5):3034, 2006.

[70] Péter Pázmány. High Resolution Speech F 0 Modification Tamás Bárdi. *Current*, pages 0–3.

[71] R. Pfeifer. Emotions in robot design. *Proceedings of 1993 2nd IEEE International Workshop on Robot and Human Communication*, pages 408–413, 1993.

[72] Dipartimento Psicologia and Università Trieste. The Music of Speech: Electrophysiological Approach. 2002.

[73] Juan G. Roederer. *The physics and psychophysics of music*. 1995.

[74] Jia Rong, Yi-Ping Phoebe Chen, Morshed Chowdhury, and Gang Li. Acoustic Features Extraction for Emotion Recognition. *6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)*, (Icis):419–424, 2007.

[75] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, 23(19):2507–17, October 2007.

[76] D Sammler, S Koelsch, T Ball, a Brandt, C E Elger, a D Friederici, M Grigutsch, H-J Huppertz, T R Knösche, J Wellmer, G Widman, and a Schulze-Bonhage. Overlap of musical and linguistic syntax processing: intracranial ERP evidence. *Annals of the New York Academy of Sciences*, 1169:494–8, July 2009.

[77] H Sato, Y Mitsukura, M Fukumi, and N Akamatsu. Emotional Speech Classification with Prosodic Prameters by Using Neural Networks. *Information Systems*, (November):18–21, 2001.

[78] K Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256, April 2003.

[79] Klaus Scherer. Which Emotions Can be Induced by Music? What Are the Underlying Mechanisms? And How Can We Measure Them? *Journal of New Music Research*, 33(3):239–251, September 2004.

[80] Klaus R. Scherer, Rainer Banse, Harald G. Wallbott, and Thomas Goldbeck. Vocal cues in emotion encoding and decoding. *Motivation and Emotion*, 15(2):123–148, June 1991.

[81] H. Schlosberg. Three dimensions of emotions. *Psychological Review*, 61:81–88, 1954.

[82] Fabian Schmieder. Suport Vector Machine in Emotion Recognition Suport Vector Machine in Emotion Recognition Author : Fabian Schmieder. *Memory*, 2009.

[83] Maartje Schreuder, Laura Van Eerten, and Dicky Gilbers. Speaking in minor and major keys.

[84] Marc Schroeder, Roddy Cowie, Ellen Douglas-cowie, Machiel Westerdijk, and Stan Gielen. Acoustic correlates of emotion dimensions in view of speech synthesis. *Emotion*, pages 1–4.

[85] Emery Schubert. Measurement and Time Series Analysis of Emotion in Music. *Emotion*, 2.

[86] Emery Schubert. Measurement and Time Series Analysis of Emotion in Music. *Emotion*, 1, 1999.

[87] David a Schwartz, Catherine Q Howe, and Dale Purves. The statistical structure of human speech sounds predicts musical universals. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 23(18):7160–8, August 2003.

[88] Robert M Seyfarth and Dorothy L Cheney. Signalers and receivers in animal communication. *Annual review of psychology*, 54:145–73, January 2003.

[89] Robert M. Seyfarth and Dorothy L. Cheney. Meaning and Emotion in Animal Vocalizations. *Annals of the New York Academy of Sciences*, 1000(1):32–55, January 2006.

[90] Marko Robnik Sikonja. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, pages 23–69, 2003.

[91] Nicholas A Smith and Mark A Schmuckler. Pitch-distributional effects on the perception of tonality. *Differentiation*, 1(1982):1–11, 2000.

[92] Nicholas a Smith and Mark a Schmuckler. The perception of tonal structure through the differentiation and organization of pitches. *Journal of experimental psychology. Human perception and performance*, 30(2):268–86, April 2004.

[93] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, August 2004.

[94] Isao Tokuda, Tobias Riede, and Juergen Neubauer. Nonlinear analysis of irregular animal vocalizations. *The Journal of the Acoustical Society of America*, 111(6):2908, 2002.

[95] Konstantinos Trohidis and George Kalliris. Multi-label classification of music into emotions. pages 325–330, 2008.

[96] J Turner. *On the Origins of Human Emotions: A Sociological Inquiry into the Evolution of Human Affect*. Stanford University Press, 2002.

[97] D Ververidis and C Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, September 2006.

[98] Dimitrios Ververidis and Constantine Kotropoulos. A State of the Art Review on Emotional Speech Databases. *Artificial Intelligence*, pages 1–11.

[99] Teija Waaramaa. *Emotions in Voice*. PhD thesis, 2009.

[100] Gregory H. Wakefield. Mathematical representation of joint time-chroma distributions. *Proceedings of SPIE*, 3807(July):637–645, 1999.

[101] Bin Yang and Marko Lugger. Emotion recognition from speech signals using new harmony features. *Signal Processing*, 90:1415–1423, 2009.

[102] Li-chiung Yang. Harmony and tension in Mandarin Chinese prosody: constraints and opportunities of lexical tones in discourse markers.

[103] Yi-hsuan Yang, Yu-ching Lin, Ya-fan Su, and Homer H Chen. A Regression Approach to Music Emotion Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):448–457, 2008.

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Stuttgart, den 03.02.2011 Mélanie Fernández Pradier